Mesures de confiance et traitement automatique de la parole

Benjamin Lecouteux

Habilitation à diriger des recherches

Spécialité : **Informatique**

Présentée par

Benjamin Lecouteux

préparée au sein du **Laboratoire d'Informatique de Grenoble**
dans **l'École Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique**

Mesures de confiance et traitement automatique de la parole

Thèse soutenue publiquement le **1er décembre 2021**
devant le jury composé de :

Nathalie Henrich-Bernardoni
Directeur de recherche CNRS Delegation Alpes (Présidente)
Martine Adda-Decker
Directeur de recherche, CNRS Ile-de-France Villejuif (Rapporteur)
Frédéric Béchet
Professeur à l'université Aix-Marseille (Rapporteur)
Jérôme Bellegarda
Chercheur HDR, Apple Cupertino USA (Rapporteur)
Yannick Estève
Professeur à l'université d'Avignon (Examinateur)
Laurent Besacier
Scientifique principal à Naver Labs Europe, Grenoble (Examinateur)

Table des matières

Chapitre 1

Introduction

1.1 Préambule

La reconnaissance automatique de la parole a été au cœur de mes recherches depuis 2005. Son interaction avec différentes disciplines liées à la parole comme la linguistique, la phonétique et l'informatique, a toujours suscité mon intérêt. En 2010 j'ai rejoint l'équipe GETALP (Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole) du LIG (Laboratoire Informatique de Grenoble) dans le cadre d'un post-doctorat. Cette année-là j'ai principalement collaboré avec Michel Vacher et François Portet sur la reconnaissance de la parole dans le cadre d'un appartement intelligent et avec Laurent Besacier qui m'a initié à la traduction automatique. En 2011, j'ai obtenu un poste de Maître de Conférence au sein de l'Université Pierre Mendés-France, intégrée depuis dans l'Université Grenoble Alpes. Je suis affecté à l'IUT2 (Institut Universitaire Technologique), département information/communication, de Grenoble pour mes enseignements et au LIG pour la recherche. J'ai ainsi rejoint officiellement l'équipe GETALP. La reconnaissance de la parole représente le fil conducteur de mes recherches que ce soit dans le cadre des appartements intelligents ou de la traduction automatique de la parole. Les évolutions dans ce domaine, tant en termes de performances que de déploiements industriels ont été considérables et nos travaux ont dû s'adapter en conséquence. C'est ainsi que, afin d'exploiter de manière plus robuste les systèmes de reconnaissance de la parole ou de traduction automatique, une partie de mes recherches s'est articulée autour des mesures de confiance. J'ai étroitement collaboré avec Michel Vacher, François Portet et Laurent Besacier au cours de la période 2010-2019 puis ensuite avec Didier Schwab. L'opportunité m'a alors été donnée de m'initier aux techniques spécifiques de la traduction automatique de la

parole vers des pictogrammes.

Avant de présenter mes travaux, réalisés cette dernière décennie autour de la parole, je souhaite rappeler brièvement les énormes mutations des systèmes de reconnaissance automatique de la parole : de quelle manière ces derniers ont évolué, comment ils se sont démocratisés et où en est la recherche en 2021. J'évoquerai ensuite la progression des mesures de confiance pour la reconnaissance automatique de la parole et présenterai l'organisation du présent manuscrit.

1.2 Un bref historique de la reconnaissance automatique de la parole

Le premier système de reconnaissance automatique de la parole a été réalisé par Bell en 1952 [Davis et al., 1952] et était capable de reconnaître des chiffres isolés prononcés par un interlocuteur spécifique. À l'époque, la technologie s'appuyait sur les formants présents dans les différents chiffres afin de les discriminer. Commencèrent ensuite à apparaître les systèmes basés sur l'alignement dynamique pour reconnaître des mots complets [Vintsiuk, 1968]. Un progrès substantiel fut l'utilisation en 1970 du spectre LPC (*Linear Predictive Coding*) permettant d'estimer le signal en se basant sur des cœfficients de prédiction [Itakura and Saito, 1970]. Anecdotiquement la reconnaissance automatique de la parole a inspiré, dès ces années 70, nombre de scénarios de films de science-fiction. Néanmoins à l'époque les capacités de calcul n'étaient pas en adéquation avec ce type de technologie. En 1976 [Reddy, 1976] prédisait que la technologie évoluerait imparablement dans la décennie à venir. Finalement, une quarantaine d'années auront été nécessaires pour atteindre des systèmes utilisables par le grand public.

L'équipe de Reddy a ainsi participé au développement de plusieurs systèmes de reconnaissance automatique de la parole : Hearsay le premier système de reconnaissance de la parole continue, Dragon qui a introduit les modèles de Markov cachés (*Hiden Markov Model* HMM) et Harpy qui a introduit la recherche par faisceau (*beam search*) , permettant ainsi d'identifier un millier de mots différents [Lowerre, 1976]. Les principales évolutions suivantes furent apportées par Sphinx 2 en 1992 qui implémentait le partage d'état au sein des HMM. Cette technologie, née dans les années 90, est ensuite devenue l'un des socles de la reconnaissance automatique de la parole jusqu'aux années 2000.

Une autre base de la reconnaissance automatique de la parole est la quantité de données d'apprentissage alliée à une forte puissance de calculs. C'est ainsi

que les modèles acoustiques ou de langage sont appris : les outils d'apprentissage machine permettent de capter indirectement les informations présentes dans ces masses de données. Dans cette optique les paramètres des HMM sont estimés via des algorithmes d'Espérance/Maximisation (EM) sur des milliers d'heures de signal. Jusqu'aux années 2010, les paramètres d'observation des HMM étaient représentés par des ensembles de Gaussiennes (*Gaussian Mixture Model* GMM), auxquels par la suite se substitueront des réseaux de neurones profonds [Hinton et al., 2012]. Les premiers systèmes dits hybrides sont alors nés : ils avaient l'avantage de profiter de toutes les avancées liées aux systèmes à base de HMM tout en modélisant mieux les observations. Nous reviendrons sur les réseaux de neurones profonds dans la prochaine section.

La modélisation du langage s'est longtemps appuyée sur des technologies qui ont peu évolué depuis les années 70 : les modèles n-grammes sont toujours utilisés dans certains systèmes de reconnaissance ; les plus grands progrès réalisés récemment sont liés aux représentations continues via des réseaux de neurones récurrents et aux représentations auto-supervisées textuelles [Devlin et al., 2019]. Mais à l'instar des modèles acoustiques, ces représentations nécessitent des quantités de données massives pour être performantes. De même la taille du vocabulaire n'a cessé d'augmenter. Les limitations étant essentiellement liées à des problèmes de mémoire et de puissance de calcul. Toutefois cet obstacle est voué à disparaître avec les approches basées sur les réseaux de neurones profonds concaténant des sous-unités de mots (*Byte Pair Encoding* BPE).

Les plus grands progrès réalisés dans les années 90 et 2000 sont liés aux techniques d'apprentissage comme les modèles indépendants du locuteur, les modèles discriminants, les modèles adaptés au locuteur etc.

Des techniques très diverses de décodage se sont développées dans le but de réduire les calculs et d'optimiser l'équation d'un système de RAP (SRAP). Ainsi, les principaux algorithmes utilisés ont été les décodeurs A*, le décodage en faisceau et les compositions de transducteurs à états finis (*Finite State Transducer* FST) : ces derniers ont été de plus en plus utilisés à partir des années 2000 en raison de l'augmentation des capacités de mémoire et de calcul.

Les corpus d'apprentissage se sont étoffés au fur et à mesure. Les premiers s'attachaient à de la lecture, puis des efforts ont porté sur la réalisation d'enregistrements (et leurs annotations) dans des conditions de plus en plus complexes. L'inconvénient majeur est que l'obtention de ces données est extrêmement coûteuse et chronophage. Aujourd'hui, seules quelques grandes sociétés ont la capacité de récolter de manière totalement non supervisée des données d'apprentissage : par exemple Google qui exploite les requêtes vocales en appliquant

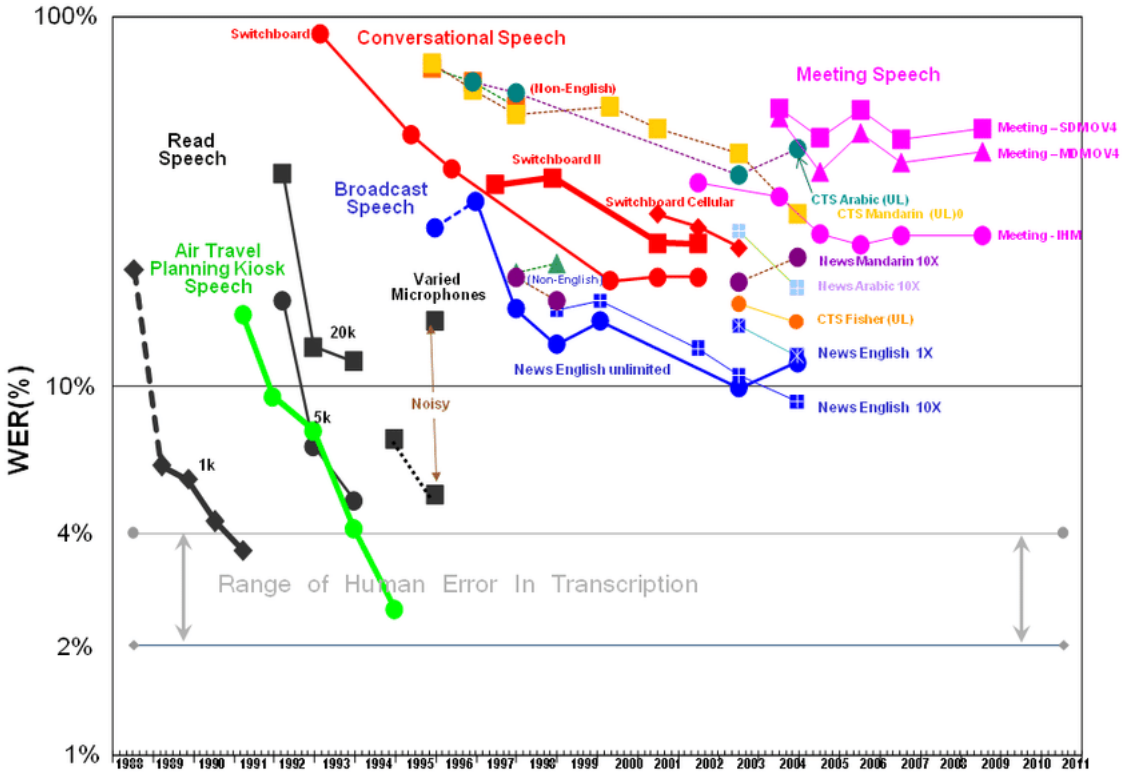


FIGURE 1.1 – Historique des résultats aux campagnes d'évaluation NIST sur différents corpus, jusqu'en 2010. Figure extraite de [van der Werff, 2012].

des heuristiques du type "si l'utilisateur consulte le lien, la reconnaissance est correcte". Par ailleurs, ces mêmes majors ont la capacité d'extraire du web des quantités de données conséquentes pour entraîner leurs modèles de langage.

La figure 1.1 montre les progrès des systèmes de reconnaissance de la parole depuis la fin des années 80 jusqu'en 2010. Plusieurs constats peuvent être faits : en fonction des tâches les résultats sont plus ou moins performants. S'ils commencent à devenir acceptables lorsqu'ils concernent les transcriptions de *broadcast news* ou la parole lue, la reconnaissance de la parole en réunions ou lors de conversations entre deux locuteurs (*switchboard*) atteint par contre des taux d'erreurs très élevés. Cependant au début des années 2010, un bouleversement a eu lieu dans la communauté parole : l'avènement des réseaux de neurones profonds.

1.3 Les avancées liées aux réseaux de neurones profonds

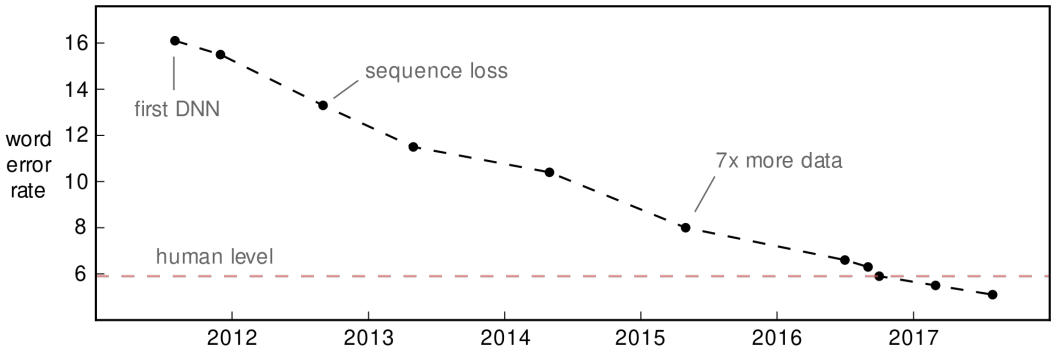


FIGURE 1.2 – Évolution des taux d'erreurs sur la tâche *switchboard* depuis le premier réseau de neurones profonds dédié à la reconnaissance automatique de la parole. Figure extraite de [Awni, 2018].

Les réseaux de neurones profonds (*Deep Neural Network* DNN) sont des structures composées de multiples couches cachées formées de neurones artificiels. Ces outils sont devenus depuis les années 2012 incontournables dans le domaine de la reconnaissance automatique de la parole et de nombreux autres domaines du TALN. Les DNN ont plusieurs avantages : ils sont plus compacts que les GMM et capables de représenter des fonctions non linéaires complexes, ce qui est très intéressant pour traiter des informations comme la parole.

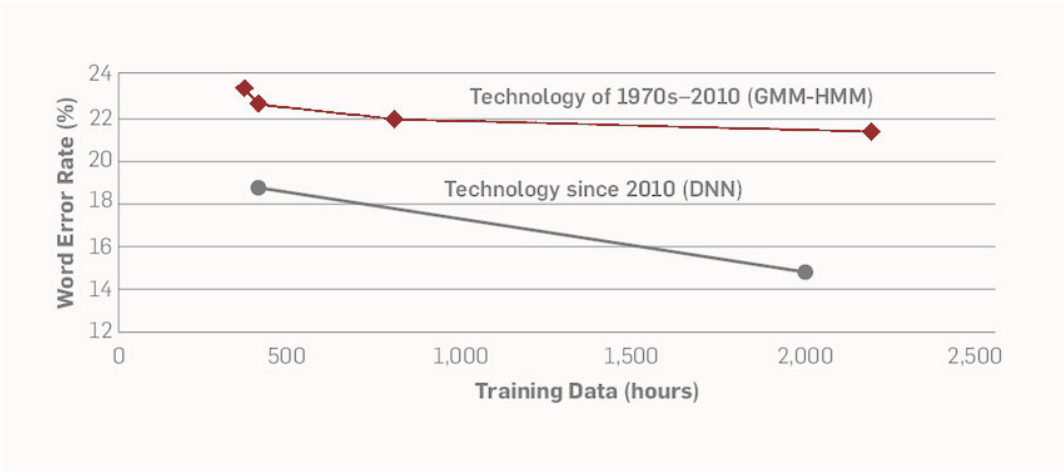


FIGURE 1.3 – Les DNN permettent de mieux tirer parti de la quantité de données. Figure extraite de [Huang et al., 2014].

Un autre avantage des architectures neuronales est qu'elles tirent un meilleur parti des grandes quantités de données, comme le montre la figure 1.3. Les systèmes état de l'art ont toujours nécessité plusieurs milliers d'heures de signal annotées pour l'entraînement de leurs modèles. Alors que les modèles GMM-HMM commençaient à atteindre un plafond de verre : doubler la quantité de données (1000 à 2000 heures) ne permettait qu'un gain relatif du taux d'erreurs mots de 2%, les modèles basés sur les réseaux de neurones profonds se sont plus montrés performants ; à quantité équivalente de données le gain relatif est de 15% [Huang et al., 2014] !

La figure 1.2 montre l'évolution des taux d'erreurs sur le corpus d'évaluation *switchboard*. Ils sont aujourd'hui de 4.9 %[1]. Le fait que les DNN soient utilisés dans de nombreux domaines autres que la parole a accéléré les recherches autour des réseaux de neurones profonds : ainsi sont apparues entre autres les architectures *sequence-to-sequence*, à base d'attention et finalement la célèbre architecture *transformer* [Vaswani et al., 2017].

Il ne faut pas perdre de vue que la reconnaissance automatique de la parole n'est pas toujours un objectif intrinsèque, mais correspond plutôt à une démarche initiale permettant de résoudre des problèmes encore plus complexes comme la compréhension, le dialogue, la traduction etc.

1.4 Défis de la reconnaissance automatique de la parole et ses limites

Les progrès évoqués ont engendré depuis 2010 une amélioration constante des performances : et depuis 2015 les systèmes sont suffisamment performants pour être utilisés dans des applications, certes encadrées, à l'échelle industrielle. La reconnaissance automatique de la parole est désormais connue du grand public et s'est déployée dans de nombreux domaines. On la retrouve ainsi sur les smartphones, les assistants (Alexa, Google Home, Siri, ...) les systèmes de guidage GPS, dans des systèmes de dialogue sur des plateformes téléphoniques, la transcription automatique de vidéos sur Youtube, le sous-titrage en temps réel de réunion dans Microsoft teams, etc. La technologie n'est donc plus du domaine de la science-fiction. Ses applications commencent à rentrer dans les mœurs.

Cependant, il existe encore des verrous concernant la transcription de parole spontanée, la parole accentuée, les conditions bruitées, multi-locuteurs etc. Les

1. https://github.com/syhw/wer_are_we

systèmes actuels ne contextualisent pas à long terme ce qu'ils transcrivent et ce manque de représentation sémantique globale est une limite. Une autre limite des systèmes est leur capacité à généraliser : ils fonctionnent très bien sur des tâches spécifiques, mais dès que les conditions s'éloignent de l'apprentissage, la dégradation des résultats est inéluctable. Ce comportement est lié à la fois à l'acoustique et à la représentation linguistique. Il en résulte que la reconnaissance automatique de la parole spontanée ou l'adaptation rapide à de nouveaux domaines restent encore un défi. À ce propos, [Likhomanenko et al., 2021] abordent la problématique d'évaluer correctement un système de reconnaissance de la parole : les auteurs mettent en avant la nécessité de multiplier les *benchmarks* pour valider une approche transférable à des enregistrements "issus du monde réel". Ils montrent également que l'ajout systématique de bruit et réverbération dans les données d'apprentissage améliore la robustesse des modèles. Il ne faut pas oublier non plus que les systèmes de reconnaissance sont performants pour un nombre restreint de langues, en l'occurrence celles où de grandes quantités de données d'apprentissage sont disponibles, mais demeurent aujourd'hui inexistants pour une majorité de langues.

Jusqu'à une date récente, la quantité de données annotées destinées à l'apprentissage a été un frein : en 2019 il était encore nécessaire d'avoir une centaine d'heures annotées pour entraîner un système "correct", des milliers d'heures pour entraîner un système état-de-l'art, voire des dizaines de milliers d'heures comme c'est le cas pour les systèmes de transcription de Google en anglais. Ces quantités de données nécessaires ont mis à mal beaucoup d'acteurs de la communauté qui ne pouvaient plus rivaliser. Toutefois les travaux portant sur l'apprentissage auto-supervisé [Schneider et al., 2019, Baevski et al., 2020] ont apporté, en partie, une réponse et permettent de surmonter cet écueil : désormais il est possible d'apprendre des représentations sans annotation sur de grandes quantités de données et quelques heures annotées suffisent ensuite pour concevoir un système de reconnaissance proche de l'état-de-l'art. Aujourd'hui (en 2021), beaucoup de recherches se tournent vers les approches auto-supervisées qui permettent de lever le grand verrou des données annotées, mais pour l'instant ces approches nécessitent le concours des super-calculateurs pour entraîner les modèles sur plusieurs milliers d'heures. Émergent alors d'autres écueils connexes : l'accès au calcul et l'impact écologique des modèles [Parcollet and Ravanelli, 2021].

1.5 Mesures de confiance et reconnaissance automatique de la parole

Les mesures de confiance, que ce soit en parole ou traduction automatique, répondent au besoin de pouvoir identifier si le système commet des erreurs ou non.

Les systèmes des années 2000 ou les systèmes de bout-en-bout, n'ont pas la capacité (ou alors très limitée) d'évaluer leur faillibilité, alors qu'*a contrario*, l'identification des erreurs est indispensable si le système n'est qu'un composant servant d'entrée pour un autre composant. Les premières approches explicites de mesures de confiance ont été proposées dans les années 90. Par exemple [Young, 1994] propose un calcul de probabilité *a posteriori* en guise de mesure de confiance. Dans les mêmes travaux, les auteurs proposent d'avoir recours à une normalisation des scores acoustiques et de combiner ceux-ci avec des connaissances de plus haut niveau liées au langage.

En 1995 [Rose et al., 1995] formalise les mesures de confiance à partir d'une étude statistique : le test du rapport de vraisemblance (*Likelihood ratio testing* (LRT)). Ainsi par la suite, des approches discriminantes à base de LRT vont naître et améliorer les performances des systèmes de RAP. Les principales mises en application vont se faire via l'interaction entre l'utilisateur et un système de dialogue qui lancera des alertes et demandera des corrections quand il l'estimera nécessaire.

Les approches automatiques les plus courantes actuellement consistent à extraire le maximum de paramètres internes et/ou externes aux systèmes et d'apprendre à un classifieur à reconnaître des indices relatifs à la mise en difficulté du système. Cependant, dans la littérature, on trouve les modèles de rejet [Sukkar et al., 1996], les anti-modèles [Sukkar et al., 1996], les cohortes [Rahim et al., 1997] qui abordent de manière différente la détection d'erreurs. Nous ne détaillerons pas ces approches car au final il s'avère [Falavigna et al., 2002] qu'elles étaient moins robustes que les probabilités *a posteriori* des mots. D'autres mesures de confiance de plus haut niveau ont été proposées comme LSA (*Latent Semantic Analysis*) [Cox and Dasmahapatra, 2002] ou l'information mutuelle inter-mots [Guo et al., 2004], mais elles se sont avérées difficiles à exploiter après décodage.

Finalement, les mesures de confiance sont intéressantes à exploiter lors d'un processus cascadant différents systèmes. Il ressort de la littérature qu'elles peuvent, en l'espèce, être utiles et améliorer la qualité des systèmes, surtout si un opérateur humain interagit avec les mesures.

Un nouveau paradigme des mesures de confiance a récemment émergé avec l'utilisation des réseaux de neurones profonds car la décision finale se fait généralement sur la couche de sortie indiquant déjà la confiance maximale du réseau. Il est alors nécessaire de mettre en place des systèmes externes qui permettent d'analyser l'état des couches ou la distribution des paramètres d'attention pour déterminer, ou non, une anomalie [Woodward et al., 2020, Fomicheva et al., 2020]. Les résultats de recherche sur les mesures de confiance pour les DNN RAP ou TA sont assez récents [Kumar and Sarawagi, 2019, Fomicheva et al., 2020, Li et al., 2021, Qiu et al., 2021, Ogawa et al., 2021]. Les travaux présentés dans ce manuscrit autour des mesures de confiance datant de la période 2011-2017 s'articuleront donc autour des mesures de confiance créées à partir de paramètres extraits.

1.6 Organisation du manuscrit

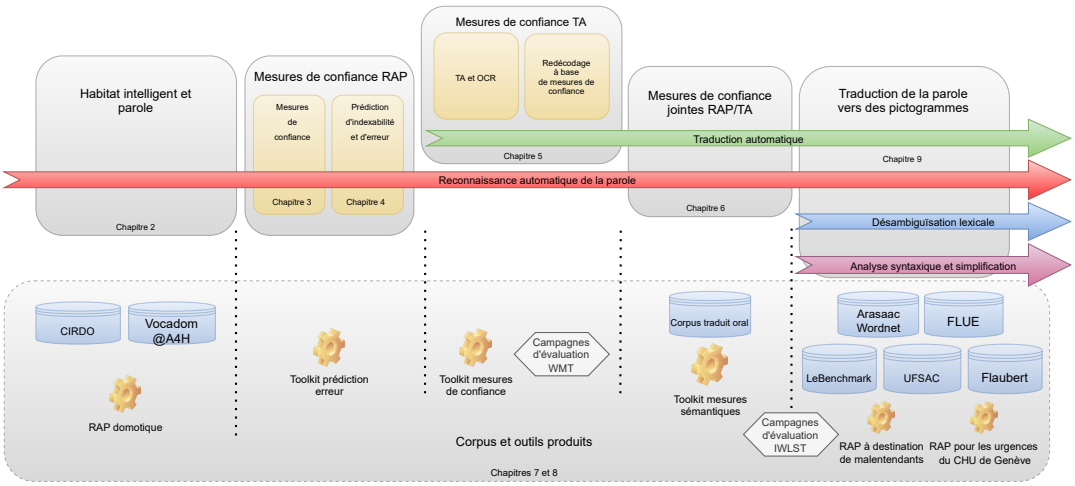


FIGURE 1.4 – Vue schématique de mes travaux autour de la RAP, TA et des mesures de confiance.

Les recherches menées ont été réalisées parallèlement aux évolutions de la reconnaissance automatique de la parole. Nous avons adapté en conséquence l'utilisation des SRAP en fonction de leurs capacités de transcription et nous avons étudié leurs applications dans les domaines de l'habitat intelligent et de la traduction automatique de la parole. Une autre évolution qui est apparue cette dernière décennie, de manière générale pour le TALN, a été la multiplication des partenariats industriels où nous effectuons des transferts technologiques de nos outils de recherche.

L'organisation du manuscrit s'articule autour de 10 chapitres, également présentés dans la figure 1.4, où j'effectue une synthèse des différents travaux auxquels j'ai contribué et se regroupant autour des thématiques suivantes :

- la parole dans l'habitat intelligent (chapitre 2).
- les mesures de confiance et la prédiction d'erreurs (chapitres 3, 4, 5, 6) pour la reconnaissance automatique de la parole et la traduction automatique de la parole.
- les campagnes d'évaluation, réalisations industrielles et corpus (chapitres 7 et 8).
- les travaux préliminaires autour de la traduction automatique de la parole vers des pictogrammes (chapitre 9).

Le chapitre 10 conclut ce manuscrit en présentant mes prochaines activités de recherche autour du projet ANR international PROPICTO récemment accepté, puis en ouvrant quelques perspectives à court et moyen terme.

Au cours du manuscrit j'alterne entre le "je" et le "nous" : le "nous" fait référence aux travaux réalisés en équipe. Pour chaque chapitre, quelques résultats illustrent les travaux. Cependant, les corpus et protocoles sur lesquels les expériences ont été réalisées ne sont pas toujours exhaustivement détaillés : le lecteur pourra se référer aux articles pointés dans chaque section.

Chapitre 2

Reconnaissance automatique de la parole et domotique

Projets : Sweet-Home (ANR 2009-2013), Vocadom (ANR 2017-2021)

Sélection de publications relatives à ce chapitre :

M. Vacher, F. Aman, S. Rossato, F. Portet, **B. Lecouteux**. Making emergency calls more accessible to older adults through a hands-free speech interface in the house. ACM Transactions on Accessible Computing 2019.

B. Lecouteux, M. Vacher, F. Portet. Distant Speech Processing for Smart Home Comparison of ASR approaches in distributed microphone network for voice command. International Journal of Speech Technology 2018.

M. Vacher, S. Bouakaz, ME. Bobillier-Chaumon, F. Aman, RA. Khan, S. Bekkadja, F. Portet, E. Guillou, S. Rossato, **B. Lecouteux**. The CIRDO Corpus : Comprehensive Audio/Video Database of Domestic Falls of Elderly People. LREC 2016.

B. Lecouteux, M. Vacher, F. Portet « Distant Speech Recognition in a Smart Home : Comparison of Several Multisource ASRs in Realistic Conditions. Interspeech 2011.

Lors de mon post-doctorat dans l'équipe GETALP, j'ai collaboré avec Michel Vacher et François Portet dans le cadre des maisons intelligentes destinées aux personnes âgées.

J'ai ainsi participé à l'élaboration des systèmes de reconnaissance automatique de la parole dans le cadre de l'habitat intelligent et plus particulièrement au sein des projets ANR Sweet-Home (2009-2012) puis Vocadom (2017-2021) portés par Michel Vacher et François Portet. En 2010, nous avons utilisé le système de reconnaissance automatique de la parole *Speeral* [Nocéra et al., 2002] développé au sein du laboratoire informatique d'Avignon et que j'avais utilisé durant ma thèse de doctorat. En 2012, nous avons migré vers *Kaldi* [Povey et al., 2011b] qui présentait une solution plus moderne et état-de-l'art. Les travaux présentés dans ce chapitre se concentrent donc sur *Kaldi*. Sweet-home a été à l'origine de nombreuses publications [Lecouteux et al., 2011b, Lecouteux et al., 2011c, Vacher et al., 2011, Vacher et al., 2012, Lecouteux et al., 2012, Sehili et al., 2012, Vacher et al., 2013a, Vacher et al., 2013b, Vacher et al., 2013c, Vacher et al., 2013d, Vacher et al., 2014a, Vacher et al., 2014b, Vacher et al., 2014c, Vacher et al., 2015a, Vacher et al., 2015b, Vacher et al., 2015c, Vacher et al., 2015d, Vacher et al., 2016a, Vacher et al., 2016b, Aman et al., 2016a, Aman et al., 2016b, Lecouteux et al., 2018, Vacher et al., 2019], dont je synthétise dans ce chapitre le contenu axé autour de la reconnaissance automatique de la parole avec *Kaldi*.

2.1 Maison intelligente et reconnaissance automatique de la parole

Domotique et maisons intelligentes

Cette dernière décennie, nous avons vu l'essor de la domotique qui donne la possibilité de piloter à distance tous les organes électriques du domicile (prises, éclairage, volets, thermostat, etc.), de monitorer son environnement (température, hygrométrie, état des ouvertures ...), de rendre en quelque sorte la maison "intelligente". La domotique représente une très prometteuse perspective en vue de contrôler et automatiser l'habitat, de le sécuriser, de le doter d'un réel confort tout en le gérant de manière plus économique. Pour ce, elle exige :

- Confort et économies énergétiques : de manager avec souplesse le chauffage ou la climatisation du logis à l'aune des variations météorologiques, de

gérer les éclairages et certains appareils électriques en fonction des besoins.

- Sécurité : d'accompagner les personnes âgées au quotidien, autorisant alors leur maintien à domicile. Elle permet de rassurer les proches (en générant des alertes en cas d'anomalie détectée par exemple) et elle peut suppléer, sans risque, à certaines tâches (fermeture de volets, gestion d'appareils ménagers, vérification des fermetures), appeler si besoin les secours, etc. Il est ressorti de nos études, dans le cadre des projets Sweet-Home et Vocadom, auprès de personnes âgées, qu'elles étaient majoritairement prêtes à investir dans leur logement afin de pouvoir y demeurer le plus longtemps possible.

Pour que l'habitat soit "intelligent", il est nécessaire qu'il réponde à un certain nombre de critères :

- Il doit être pourvu de capteurs permettant de récolter des informations telles que les mouvements, les états des ouvertures, la température etc.
- il doit être équipé de manière à automatiser au mieux les actions de la vie quotidienne comme la fermeture des volets, le contrôle de l'éclairage et des appareils ménagers (TV, radio, aspirateur).
- il doit au final disposer d'une entité capable d'analyser les informations remontées par les capteurs et en déduire des actions compatibles avec le profil des utilisateurs.

Les projets ANR Sweet-Home et Vocadom se sont prioritairement attachés au maintien à domicile des personnes, lequel implique de mettre en place des fonctionnalités permettant de veiller sur la personne et faire appel à une assistance dans des situations anormales, de l'aider, autant que nécessaire, à compenser une dépendance. Plusieurs freins ou difficultés à la démocratisation des solutions de maintien à domicile ont été identifiés :

- le respect de la vie privée : les données récoltées sont très personnelles et renseignent sur l'activité de la personne, sur ce qu'elle dit s'il y a des micros etc. Par exemple, certaines personnes nous ont confié qu'elles avaient peur que des informations extraites de ces systèmes ne soient un argument pour des proches en vue de les placer en institution.
- l'adaptation au profil de l'utilisateur : les habitudes diffèrent en fonction de la personne, ses attentes, ses dépendances. Les solutions doivent être flexibles et simples à utiliser.
- la multiplicité des systèmes et protocoles domotiques qui ne sont pas normalisés et se révèlent souvent incompatibles entre eux. Aujourd'hui chaque constructeur essaye de faire émerger ses propres solutions et

protocoles. Il faut cependant souligner que plusieurs projets tendent à assurer un échange universel entre toutes les solutions existantes : homeassistant, openhab, jeedom, domoticz ... Au niveau des protocoles, un consensus commence à se dessiner autour du protocole ouvert *Matter*, qui nécessite une certification dans le cadre d'une utilisation commerciale.

La première préoccupation de Sweet-Home et Vocadom s'est focalisée comme nous venons de l'exposer, sur le respect de la vie privée des personnes : les études ont montré que l'acceptation de solutions intelligentes était plus large en limitant la sortie d'informations du domicile. Pour ce faire il est nécessaire de traiter les données *in situ*. La seconde préoccupation a été le choix de l'interface homme-machine. Il s'est porté sur l'utilisation d'une modalité qui était encore peu utilisée au moment de l'initiation des projets (2009) : l'usage de la parole. Si ce type d'interface commençait à se répandre dans les systèmes de terminaux mobiles [Bellegarda, 2013], son usage dans les systèmes domotiques du commerce n'est apparu que récemment. La principale raison résidant dans les difficultés de reconnaissance de la parole en conditions dégradées : bruit, distance, etc. L'usage de la parole apporte plusieurs avantages : c'est un moyen de communiquer naturel et qui permet de détecter des situations d'appel à l'aide.

RAP dans l'habitat

L'utilisation de la parole pour "commander" l'habitat nous est ainsi apparue comme le moyen le plus naturel de communication au sein de l'habitation : un système idéal communicant avec la parole permet de s'abstraire de la connaissance d'outils numériques (écran de contrôle, télécommande etc.) et s'adresse ainsi à un public plus large. Ce type d'approche a d'ailleurs été confirmé ces dernières années avec l'intégration d'outils à base de commande vocale grand public dans l'habitat. Cependant, d'autres aspects sont également à prendre en considération : les coûts des solutions, leur impact social et les aspects liés à la vie privée. Dans le cadre des projets ANR Sweet-Home et Vocadom, nous estimons, par exemple, que la reconnaissance automatique de la parole doit être réalisée en local et ne pas dépendre de services extérieurs. Ceci permet de résoudre à la fois des problèmes liés à la vie privée (maîtrise des données personnelles) et à la sécurité (le système doit rester autonome en cas de coupure internet par exemple). Une autre volonté des projets est d'utiliser du matériel courant à faible coût (que ce soient les ordinateurs, les capteurs, les micros etc.). Une attention particulière a été prêtée à l'impact social pour trouver le positionnement idéal de l'utilisation de nos systèmes : ne pas les

mettre au service d'une surveillance orwellienne, ne pas réduire l'autonomie des personnes par une assistance disproportionnée etc.

Dans le contexte de l'habitat, la reconnaissance de la parole classique est en difficulté en raison des spécificités de la voix et de l'environnement distant. Concernant les aspects liés à la sécurité de la personne, nous avons souhaité être en mesure de détecter des appels à l'aide (personne qui est tombée) ou des bruits anormaux (une chute). Une nouvelle entrave à la reconnaissance tient aux paroles qui sont prononcées en situation de détresse : l'émotion altère l'audibilité des paroles et les rend plus complexes à interpréter.

Les projets Sweet-Home et Vocadom sont extrêmement conséquents, ils ont abordé les études d'usage, la détection de bruits, la détection de locuteurs, la scénarisation en fonction des informations remontées par les capteurs, la constitution de corpus écologiques. Dans ce chapitre, je n'aborde que les aspects liés à la reconnaissance automatique de la parole auxquels j'ai principalement contribué.

Les principaux défis ont ainsi été la mise en place de systèmes RAP *in situ* capables de reconnaître une série d'ordres domotiques, des appels de détresse et le tout en conditions distantes avec des micros placés aux plafonds. Une autre difficulté a été la quantité de données relativement faible en conditions réelles pour adapter les modèles acoustiques.

2.2 Recherches dans le cadre du projet ANR Sweet-Home

La perspective de pouvoir commander l'habitat intelligent à travers une interface simple, à savoir la voix, est assurément attrayante. Mais elle nécessite de surmonter de réelles difficultés. La mise en œuvre de la reconnaissance vocale associée à cette démarche implique de maîtriser des enregistrements à distance des locuteurs, dans un environnement potentiellement bruité. Différentes études ont montré [Aman et al., 2013, Vacher et al., 2015a, Aman et al., 2016a] qu'il faut prendre en considération la réverbération, le bruit, l'orientation des micros etc. Une particularité du projet Sweet-Home a été la mise à disposition d'un appartement réel reconstitué : les conditions d'enregistrement sont proches de la réalité avec des micros placés dans les plafonds. La figure 2.1 présente le plan de l'appartement Domus où s'est déroulé l'ensemble des expériences en conditions écologiques et liées à la domotique.

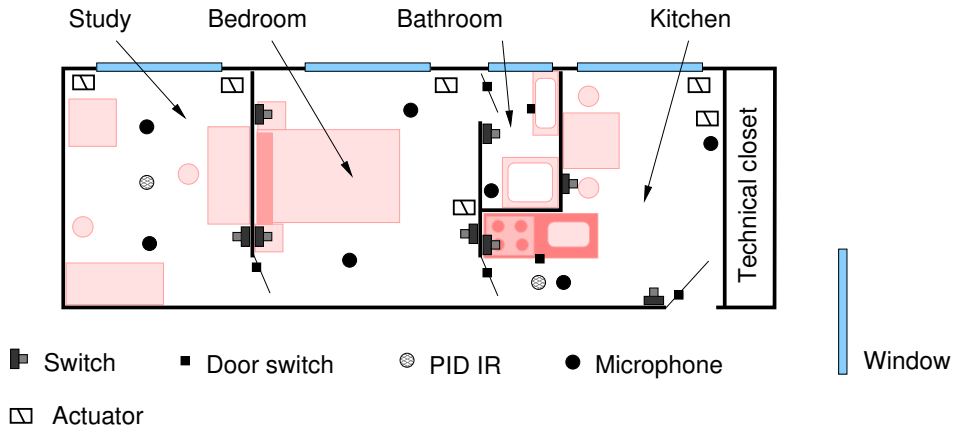


FIGURE 2.1 – Plan de Domus extrait de [Vacher et al., 2015a], l'appartement témoin utilisé pour réaliser les expériences de domotique. Les points noirs représentent les micros placés dans les plafonds.

2.2.1 Fusion dynamique de canaux de SRAP

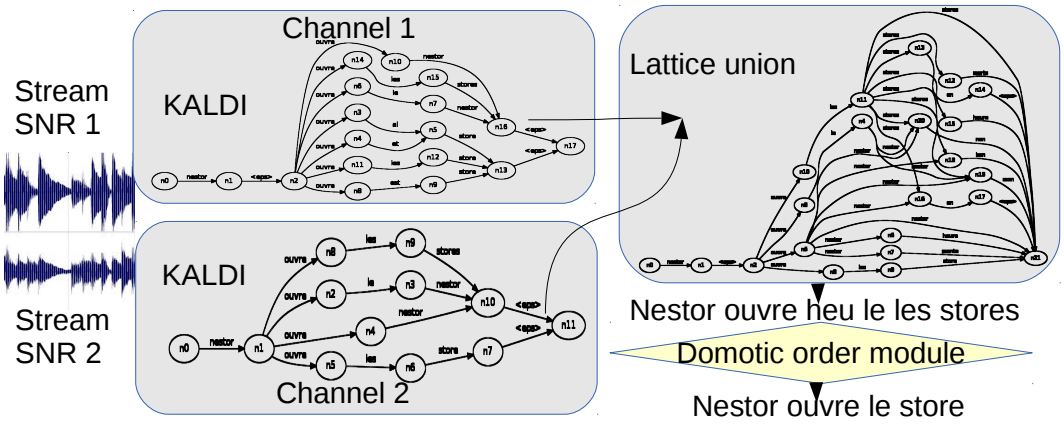


FIGURE 2.2 – Illustration [Lecouteux et al., 2018] de la fusion d'informations issues de différents micros.

Dans le cadre de Sweet-Home, je me suis particulièrement intéressé aux aspects de la RAP en conditions défavorables. La faible qualité des signaux acoustiques (distance, réverbérations, bruits ambiants etc.) représentait initialement la principale difficulté à surmonter. J'ai mené mes recherches sur la possibilité de fusionner des informations issues des différents micros positionnés dans les pièces de l'appartement expérimental. Nous avons ainsi proposé plusieurs méthodes pour mutualiser l'information captée par chacun d'entre eux. Une première voie d'investigation proposée fut d'étendre mes précédents travaux portant sur le décodage guidé

[Lecouteux et al., 2006b, Lecouteux et al., 2007, Lecouteux et al., 2013] : le SRAP génère des hypothèses au fur et à mesure de l'exploration du graphe. La meilleure hypothèse à un temps t est étendue en fonction de la probabilité de l'hypothèse courante, du résultat de l'exploration courante du graphe et des hypothèses issues d'autres SRAP (décodant le même signal, mais capté par d'autres micros). Afin de combiner l'information issue d'une transcription auxiliaire H_{aux} avec le processus de recherche, un point de synchronisation doit être trouvé pour chaque mot que le système évalue. Ces points sont trouvés en alignant dynamiquement la transcription fournie avec l'hypothèse courante ; cette tâche est effectuée en minimisant la distance d'édition entre les deux hypothèses. Ce processus permet d'identifier dans la transcription auxiliaire H_{aux}, la meilleure sous-séquence qui correspond à l'hypothèse courante H_{cur}. Cette sous-séquence H_{aux} est utilisée pour une ré-estimation du score linguistique en fonction de mesures de confiance $\phi(w_i)$ associées au système auxiliaire :

$$\begin{aligned} &L(w_i|w_{i-2}, w_{i-1}) = P(w_i|w_{i-2}, w_{i-1})^{1-\beta} \times \phi(w_i)^{\beta} \\ &\beta = 0 \text{ si } w_i \text{ n'est pas trouvé dans } H_{aux} \end{aligned} \tag{2.1}$$

Où $L(w_i|w_{i-2}, w_{i-1})$ est le score linguistique résultant, $P(w_i|w_{i-2}, w_{i-1})$ est la probabilité initiale du tri-gramme, β est un facteur d'échelle estimé empiriquement et $\phi(w_i)$ correspond à un score de confiance associé au mot w_i. Dans cette situation, l'idée est de mutualiser les informations explorées par différents systèmes de reconnaissance automatique de la parole. Le décodage guidé tel que décrit a été utilisé courant 2010 avec le SRAP *Speeral*.

Par la suite, nous avons proposé de généraliser le décodage guidé en fusionnant les espaces de recherche, via l'union des graphes FST issus de différents SRAP, sur différents canaux afin de faire émerger un consensus. Cette approche, a permis d'obtenir des améliorations significatives en terme de détection de commandes vocales. En effet, chaque micro perçoit une variante de l'ordre domotique proposé et les informations issues de chacun s'avèrent complémentaires. La figure 2.2 illustre la fusion de deux canaux qui ont été décodés par deux systèmes de RAP. L'inconvénient de cette approche est qu'elle est relativement consommatrice en termes de calculs car un décodage doit être effectué pour chaque micro. Cependant, l'espace de recherche est réduit car contraint par une grammaire : cette solution reste viable.

2.2.2 Modélisation acoustique spécialisée : s-GMM

Une autre partie de nos recherches s'est axée sur l'utilisation des *subspace*-GMM : ces derniers ont permis de pallier la petite quantité de données annotées disponibles pour réaliser l'apprentissage de nos modèles acoustiques.

Les GMM et le s-GMM permettent de modéliser la probabilité d'émission de chaque état HMM avec un mélange de gaussiennes. Cependant, dans l'approche s-GMM, les gaussiennes et leurs poids sont générés à partir d'un modèle du monde UBM (*Universal Background Model*) appris sur de grandes quantités de données qui n'ont pas la nécessité d'être annotées.

Le modèle s-GMM est décrit ainsi dans [Povey et al., 2011a] :

$$\begin{cases} p(\mathbf{x}|j) = \sum\limits_{m=1}^{M_j} c_{jm} \sum\limits_{i=1}^{I} w_{jmi} \mathcal{N}(\mathbf{x}; \mu_{jmi}, \mathbf{\Sigma}_i), \\ \mu_{jmi} = \mathbf{M}_i \mathbf{v}_{jm}, \\ w_{jmi} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_{jm}}{\sum_{i'=1}^{I} \exp \mathbf{w}_{i'}^T \mathbf{v}_{jm}}. \end{cases} \tag{2.2}$$

où $\mathbf{x}$ désigne le vecteur acoustique, $j \in \{1..J\}$ est l'état HMM, i est l'indice gaussien, m est un sous-état et c_{jm} est le poids du sous-état. Chaque état j est associé à un vecteur $\mathbf{v}_{jm} \in \mathbb{R}^S$ (S est la dimension du sous-espace phonétique) qui dérive les moyennes, μ_{jmi} et les poids du mélange w_{jmi} et il a un nombre partagé de gaussiennes, I. Le sous-espace phonétique $\mathbf{M}_i$, les projections de poids $\mathbf{w}_i^T$, les matrices de covariance $\mathbf{\Sigma}_i$ i et les paramètres globalement partagés ; $\mathbf{\Phi}_i = \{\mathbf{M}_i, \mathbf{w}_i^T, \mathbf{\Sigma}_i\}$ sont communs à tous les états. Ces paramètres peuvent être partagés et estimés sur plusieurs conditions d'enregistrement. Le mélange générique de I gaussiennes, appelé modèle du monde (UBM), modélise toutes les données d'entraînement de la parole pour l'initialisation du s-GMM.

Nos expériences visent à obtenir les paramètres partagés du s-GMM en utilisant à la fois les données d'apprentissage Sweet-Home (7h) et les données propres (ESTER+REPERE 500h). En ce qui concerne la partie GMM, les trois ensembles de données d'entraînement sont simplement fusionnés en un seul. Nous proposons d'entraîner un système s-GMM classique en utilisant toutes les données que nous avions et non annotées pour entraîner l'UBM (1000 gaussiennes) : l'objectif était de biaiser spécifiquement le modèle acoustique avec les conditions de la maison intelligente et de la parole expressive.

Ainsi les s-GMM ont permis d'exploiter via le modèle UBM les données non-annotées et d'obtenir un système de reconnaissance relativement performant à partir d'une petite quantité de données annotées. Dans ces travaux, les

s-GMM ont même montré une supériorité par rapport aux modèles à base d'apprentissage profond de l'époque.

2.2.3 Fusion de canaux bas niveau : le *beamforming*

Nous avons également proposé des méthodes inspirées du *beamforming*, initialement réservé aux antennes de micros. Dans notre situation, nous avons considéré l'ensemble des micros de l'appartement comme une antenne et nous avons fusionné directement au niveau du signal l'ensemble des micros. Cette fusion se fait en deux étapes : la première consiste à sélectionner le micro "maître", qui a le signal le plus élevé et ce dernier va servir de référence pour synchroniser le signal des autres micros. Les différents signaux sont alors alignés via un alignement dynamique afin de renforcer l'information commune (la parole dans notre cas).

L'algorithme de *beamforming* se base sur la somme pondérée d'une antenne de micros. Étant donnés M microphones, le signal de sortie $y[t]$ est calculé par :

$$y[t] = \sum_{m=1}^{M} W_m[t] x_m[t - D^{(m,ref)}[t]] \tag{2.3}$$

Où $W_m[t]$ est le poids pour le micro m au temps t, avec $\sum_{m=1}^{M} W_m[t] = 1$, le signal du m^e canal est $x_m[t]$ et $D^{(m,ref)}[t]$ correspond au délai entre le m^e canal et le canal de référence. Dans nos expériences, le canal de référence est celui présentant le ratio signal/bruit le plus élevé. Les signaux sont fusionnés vers un signal mono-source y qui sera utilisé comme source unique pour la RAP.

2.2.4 Grammaires et mots clés

D'autres travaux ont porté sur les méthodes établissant la reconnaissance d'ordres domotiques. Il ressort des résultats qu'il était indispensable d'avoir recours à des grammaires représentant les différentes commandes. Cependant, il ne faut pas que le système soit trop contraint par cette grammaire : s'il y a des erreurs ou que l'utilisateur s'éloigne un tant soit peu de la grammaire, la commande vocale ne peut être reconnue. La solution optimale a donc été de mélanger des modèles de langue classique avec la grammaire afin de lui donner un peu de flexibilité. Il est également apparu un point important qui peut sembler trivial : la nécessité d'un mot clé déclenchant, lorsque cela est possible, la reconnaissance de la parole. Cependant, dans le cadre de Sweet-Home notre objectif a consisté aussi à détecter des situations de détresse. Dans cette configuration, l'hypothèse que l'utilisateur ne fera pas forcément l'usage

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

FIGURE 2.3 – Exemple d'une grammaire de commandes vocales présentée dans [Lecouteux et al., 2018].

correct d'un mot clé nous a conduit à des solutions se basant sur des mots clés pour les ordres domotiques classiques et à une reconnaissance de la parole plus ouverte en ce qui concerne les situations de détresse.

2.2.5 Détection des commandes vocales

Nous avons proposé de transcrire chaque commande vocale et sortie RAP en un graphe de phonèmes dans lequel chaque chemin correspond à une variante de prononciation. Pour chaque sortie RAP phonétisée T, chaque commande vocale H est alignée sur T en utilisant une distance de Levenshtein. Les coûts de suppression, d'insertion et de substitution ont été calculés de manière empirique tandis que la distance cumulative $\gamma(i, j)$ entre H_j et T_i est donnée par l'équation : 2.4.

$$\gamma(i, j) = d(T_i, H_j) + \min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\} \tag{2.4}$$

La fonction de distance $d()$ est biaisée en fonction de la probabilité de confusion des phonèmes.

La commande vocale ayant le meilleur score de symboles alignés est sélectionnée suivant un seuil de détection. Cette approche prend en compte certaines erreurs de reconnaissance comme les terminaisons de mots ou des variations de diction. De plus, dans de nombreux cas, un mot mal décodé est phonétiquement proche du bon mot en raison de sa prononciation proche.

Quelques résultats

La table 2.1 présente les résultats des expériences sur les corpus "interaction" (21mn, 803 ordres) et "user specific" (17 mn, 549 ordres) en utilisant différentes combinaisons de canaux et de modèles acoustiques. Nous utilisons 3 modèles acoustiques : le premier basé sur une approche HMM-GMM avec adaptation fMLLR+SAT, le second basé sur des s-GMM et enfin des modèles hybrides HMM-DNN. Nous comparons 5 méthodes :

- la sélection du meilleur canal en terme d'énergie : SNR1
- la sélection du second canal en terme d'énergie : SNR2
- fusionner les signaux des 7 micros via un algorithme de type *beamforming*
- un ROVER basé sur 7 canaux
- la fusion des espaces de recherche de SNR1 et2 : SNR1&2

Les résultats sont présentés en termes de WER (*Word Error Rate*) et DER (*Domotic Error Rate*) qui est défini ainsi :

$$DER = \frac{commandes\ oubliees + fausses\ alarmes}{cammandes\ detectees\ correctement}$$

Les meilleures performances sont obtenues avec le modèle hybride HMM-DNN ou le modèle s-GMM en fonction du corpus associés à la fusion de l'espace de recherche des deux meilleurs canaux. Le corpus *Interaction* présente des conditions d'enregistrement plus difficiles que celles du corpus *User specific* et le modèle s-GMM semble alors plus robuste. Par ailleurs, les DER montrent que les ordres domotiques sont bien reconnus malgré un WER qui reste élevé, aux alentours des 30%.

2.3 Conclusion

Dans ce chapitre, nous avons synthétisé les différentes recherches menées autour de la reconnaissance automatique de la parole dans le cadre de l'habitat intelligent. Nous avons proposé plusieurs méthodes permettant d'avoir un SRAP à la fois robuste et qui fonctionne localement.

Suite au projet Sweet-Home (fini en 2014), Michel Vacher a déposé le projet Vocadom (démarré en 2017) plus focalisé sur les personnes âgées ou en situation de handicap. Une autre évolution au sein de Vocadom a été l'utilisation d'antennes de micros au lieu de micros classiques : l'objectif étant de mettre en

	GMM-HMM + fMLLR+SAT		SGMM-HMM		DNN-HMM	
WER/DER	DEV	TEST	DEV	TEST	DEV	TEST
WER Interaction SNR1	35.00	30.95	30.65	**27.86**	29.78	29.11
WER Interaction SNR2	30.65	34.21	26.52	31.75	27.83	**31.31**
WER Interaction Beamforming	29.60	30.10	26.20	29.12	27.83	**28.49**
WER Interaction ROVER	28.70	30.10	26.33	**28.22**	27.83	28.49
WER Interaction SNR1&2	27.61	29.83	26.04	**27.22**	26.83	27.49
WER User specific SNR1	28.16	39.79	29.99	**34.88**	28.55	37.95
WER User specific SNR2	36.77	39.96	38.98	**38.34**	35.98	39.01
WER User specific Beamiforming	27.95	38.63	30.10	**35.42**	26.01	37.31
WER User specific ROVER	27.57	38.57	28.90	**35.00**	28.21	37.51
WER User specific SNR1&2	27.47	37.56	28.94	**34.38**	27.81	37.11
DER Interaction SNR1	6.98	4.75	5.43	**3.86**	6.20	5.93
DER Interaction SNR2	6.20	5.93	3.10	5.79	5.43	**5.49**
DER Interaction Beamforming	6.20	5.00	3.10	**4.56**	5.43	5.64
DER Interaction ROVER	5.43	5.00	3.30	**3.86**	5.43	5.64
DER Interaction SNR1&2	3.88	4.15	2.65	**3.86**	3.88	5.03
DER User specific SNR1	2.19	2.19	1.09	**1.91**	1.09	**1.91**
DER User specific SNR2	4.92	4.10	3.83	3.28	3.28	**2.46**
DER User specific Beamforming	2.00	2.10	1.29	1.91	2.00	**1.78**
DER User specific ROVER	1.8	2.19	1.09	1.91	1.75	**1.78**
DER User specific SNR1&2	1.64	1.37	1.09	1.73	1.09	**1.37**

TABLE 2.1 – Résultats extraits de [Lecouteux et al., 2018] : DER et WER sur les corpus *Interaction* (21 mn, 803 ordres) et *user specific* (17 mn, 549 ordres) en utilisant différentes combinaisons de canaux et de modèles acoustiques. Nous utilisons 3 modèles acoustiques : (fMLLR+SAT), (s-GMM) et (DNN). Nous comparons 5 méthodes : le meilleur canal SNR1, le deuxième meilleur canal SNR2, un *beamforming* basé sur 7 canaux, un ROVER basé sur 7 canaux et SNR1&2 combine les graphes des canaux SNR et SNR2.

place des techniques de séparation de source. Au sein de ce projet j'ai continué à travailler sur les aspects parole et nous avons ainsi proposé des solutions pour une reconnaissance automatique de la parole robuste dans l'habitat intelligent.

Les travaux présentés ont permis d'aboutir à un système complet avec détection de la parole (*Voice Activity Detection* VAD) et fonctionnant à faible latence. Nous avons également produit des corpus destinés à la RAP en habitat intelligent qui sont présentés dans le chapitre 8.

Depuis la fin du projet Sweet-Home en 2014, de grandes évolutions ont eu lieu dans la communauté de la reconnaissance automatique de la parole : le développement des techniques à base d'apprentissage profond se sont substituées aux approches précédentes. En effet, l'ajout de données synthétiques bruitées lors de l'apprentissage associées à une étape de *finetuning* sur les données cibles permettent de pallier en partie la faiblesse des quantités de données d'apprentissage. L'émergence des modèles auto-supervisés sera également un facteur d'amélioration non négligeable dans les années à venir.

Toutefois, lorsque les données ne sont pas libres ou lorsque les conditions de reconnaissance de la parole sont très particulières, des approches hybrides et à base de grammaires demeurent pertinentes.

Il est intéressant de noter une évolution assez rapide dans le domaine de la parole liée à la domotique sur la période 2009-2021. En effet, les "boxes domotiques" ou les "assistants" ont commencé à se démocratiser au sein du grand public depuis 2015 : presque chaque grand major de l'informatique a développé ses solutions. Dans la majorité des cas, se posent des questions de confidentialité [Schönherr et al., 2020] car les données liées à la parole sont souvent transmises sur des serveurs externes. Les solutions locales sont rarement d'actualité et les normes/protocoles en domotique manquent encore de normalisation pour produire des solutions fiables et simples à l'échelle industrielle.

Chapitre 3

Mesures de confiance et mots hors-vocabulaire

Projets : TRIDAN (DUAL RAPID DGA)

Encadrements : Kamel Bouzidi, Zied Elloumi

Sélection de publications relatives à ce chapitre :

K. Bouzidi, Z. Elloumi, L. Besacier, **B. Lecouteux**, MF. Benzeghiba. Traitement des Mots Hors Vocabulaire pour la Traduction Automatique de Document OCRisés en Arabe. TALN 2017.

B. Lecouteux, P. Nocera, G. Linarès (2010). « Semantic cache model driven speech recognition. ICASSP 2010.

G. Senay, G. Linarès, **B. Lecouteux**, S. Oger, T. Michel (2010). « Transcriber driving strategies for transcription aid system. LREC 2010.

B. Lecouteux, G. Linarès, B. Favre. Combined low level and high level features for Out-Of-Vocabulary Word detection. Interspeech 2009.

Les mesures de confiance - qui s'attachent à l'estimation de la qualité de la sortie d'un système, connaissant l'entrée correspondante - ont été au cœur de mes recherches depuis 2009. Elles sont des outils clés dans la reconnaissance automatique de la parole ou encore la traduction automatique. En effet, ces dernières permettent d'orienter l'utilisateur sur les zones d'incertitude. Les mesures de confiance peuvent être utilisées à différentes granularités :

- Au niveau du mot où l'on lui associe une probabilité d'être correct ou non
- Le segment : la notion de segment n'est pas forcément arrêtée, mais peut correspondre à des unités utilisées par les systèmes comme le n-gramme, le *chunk* dans le cadre de la traduction automatique etc.
- Au niveau de la phrase, où l'on donne une qualité globale à une phrase émise.
- Et plus globalement au niveau d'un document complet : ce cas d'utilisation est plus rare car très général.

Les mesures de confiance ont connu un essor dans les années 2000 ; notamment attachées aux tâches de traduction automatique ou de transcription automatique. Elles sont même devenues un outil d'assistance au post-éditeur, en l'orientant rapidement sur les zones problématiques. Dans ce chapitre nous évoquerons les travaux réalisés autour des mesures de confiance et de ses applications dans les domaine de la parole ou de la reconnaissance automatique de caractères. Dans un premier temps, nous aborderons les corrections assistées par l'humain [Senay et al., 2010b] puis nous discuterons de l'exploitation automatique des mesures de confiance pour guider un SRAP [Lecouteux et al., 2010] ou un système d'OCR (*Optical Character Recognition*) [Bouzidi et al., 2017].

3.1 Détection de mots hors-vocabulaire à partir des mesures de confiance

Mes premiers travaux autour des mesures de confiance ont débuté en 2009. Nous proposions des mesures destinées à la reconnaissance automatique de la parole, particulièrement spécialisées dans la détection des mots hors-vocabulaire (MHV) [Lecouteux et al., 2009a] :

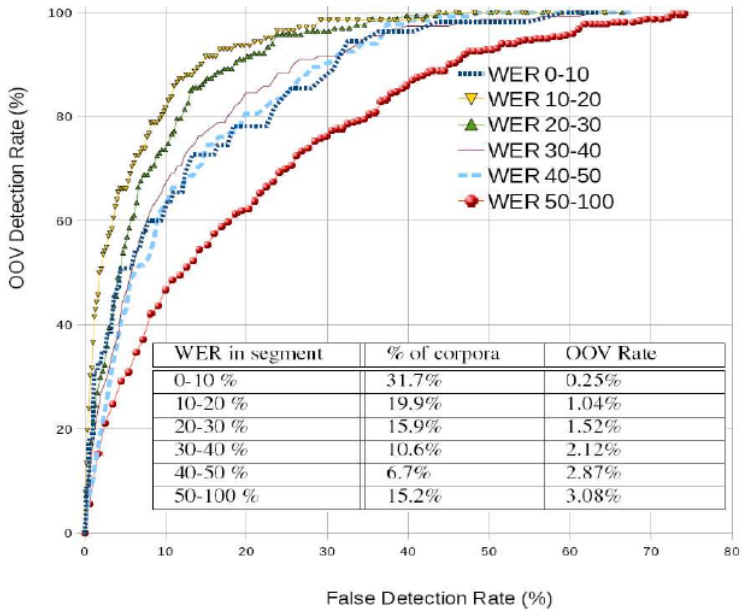
Ces derniers induisent des distorsions entre l'hypothèse et la meilleure séquence phonétique. Leur détection peut se révéler importante. Nous avons proposé des mesures de confiance permettant d'extraire le maximum d'informations issues d'un système de reconnaissance automatique de la parole en deux

étapes. La première extrait des paramètres de bas niveau relatifs à l'acoustique et à la topologie du graphe, ainsi que des paramètres de plus haut niveau liés à la linguistique. La deuxième s'opère à partir de ces paramètres ; un classifieur assigne à chaque mot une probabilité d'être correct, comme détaillé dans [Moreno et al., 2001]. Chaque mot de l'hypothèse est au final représenté par un vecteur de paramètres, qui se regroupent en 3 classes :

1. Des **paramètres acoustiques** comme la log-vraisemblance acoustique du mot et celle de la trame, ainsi que la différence entre la log-vraisemblance du mot et celle du segment correspondant en supprimant toute contrainte linguistique.
2. Des **paramètres linguistiques** basés sur les probabilités issues d'un modèle de langage utilisé dans le système de reconnaissance. Nous utilisons la probabilité linguistique, la perplexité du mot dans un espace défini. Nous adjoignons également une mesure issue de [Mauclair et al., 2006], relative au repli du mot au niveau du modèle de langage.
3. Des **paramètres liés au graphe** basés sur l'analyse du réseau de confusion. L'utilisation de ces paramètres est motivée par le fait que l'algorithme d'exploration génère de nombreux chemins alternatifs lorsqu'un mot hors-vocabulaire apparaît. Le comportement de l'exploration semble être un bon indicateur pour la détection des mots hors-vocabulaire : nous utilisons le nombre de chemins alternatifs et la probabilité *a posteriori*. Nous incluons aussi des valeurs liées à la distribution des probabilités *a posteriori* dans le réseau de confusion : maximum, minimum et moyenne des probabilités *a posteriori*.

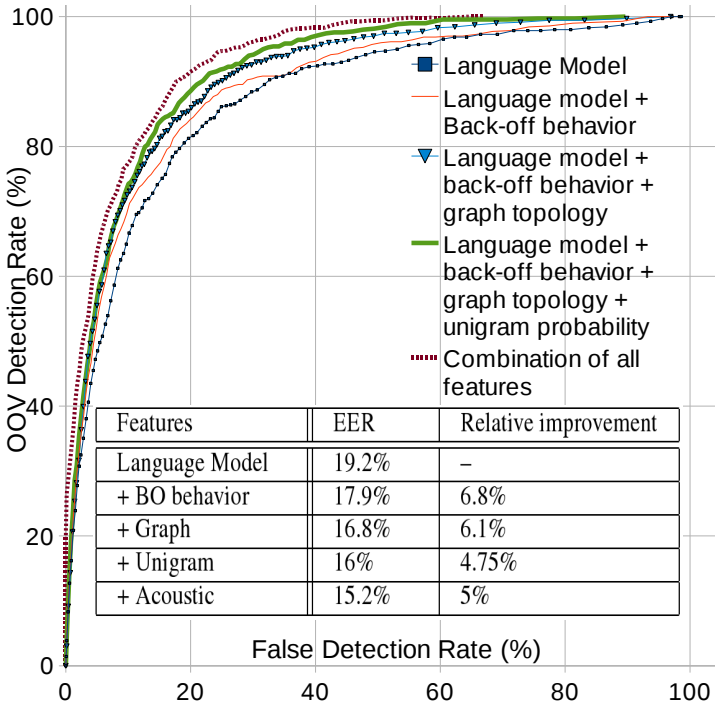
Résultats

Un algorithme de classification de type boosting permet de combiner les paramètres [Moreno et al., 2001]. Cet algorithme réalise une recherche exhaustive de la combinaison linéaire d'un ensemble de classifieurs en s'appuyant sur la sur-pondération d'exemples mal estimés. Les résultats issus de cette classification permettent ainsi d'obtenir l'une des deux classes pour chaque mot : MHV ou non-MHV, avec un score associé. La figure 3.2 présente les résultats extraits de [Lecouteux et al., 2009a, Lecouteux et al., 2009b] et portant sur 7h de test issues d'ESTER [Galliano et al., 2005] où la combinaison de différents traits permet d'identifier les MHV. L'apprentissage du modèle de détection a été réalisé sur le corpus d'apprentissage d'ESTER (200h). Nous observons que les MHV sont mieux détectés dans les tranches de performance



WER in segment	% of corpora	OOV Rate
0-10 %	31.7%	0.25%
10-20 %	19.9%	1.04%
20-30 %	15.9%	1.52%
30-40 %	10.6%	2.12%
40-50 %	6.7%	2.87%
50-100 %	15.2%	3.08%

FIGURE 3.1 – Résultats de détection des MHV en fonction des taux d'erreurs mots.



Features	EER	Relative improvement
Language Model	19.2%	–
+ BO behavior	17.9%	6.8%
+ Graph	16.8%	6.1%
+ Unigram	16%	4.75%
+ Acoustic	15.2%	5%

FIGURE 3.2 – Paramètres utilisés pour la détection de mots hors-vocabulaire.Résultats extraits de [Lecouteux et al., 2009a, Lecouteux et al., 2009b].

moyenne du système de reconnaissance, comme le montre la figure 3.1 : quand les taux d'erreur sont très élevés ou très faibles, la détection est un peu moins efficiente. Cependant, ils peuvent être détectés de manière relativement fiable et l'ensemble des paramètres acoustiques, linguistiques et graphiques s'avèrent complémentaires. Le EER (*Equal Error Rate*) présenté dans la figure 3.2 correspond au moment où le taux de fausse détection atteint un point d'équilibre avec le taux de rappel.

Ces mesures de confiances seront ré-utilisées dans les travaux présentés dans les sections 3.2 et 3.3.

3.2 Décodage interactif à l'aide de mesures de confiance

Par la suite, nous avons exploité ces mesures de confiance dans le cadre de décodage interactif de la parole ; l'idée étant d'assister l'utilisateur à la post-édition. Le paragraphe suivant reprend les travaux présentés dans [Senay et al., 2010a, Senay et al., 2010b].

Généralement, la meilleure hypothèse de reconnaissance ne constitue qu'une partie de l'information des données à transcrire appréhendées par le le système de RAP. Proposer des alternatives au correcteur permet d'améliorer son efficacité [Nanjo et al., 2006], par exemple en lui évitant la saisie de certains mots. Cette approche présente néanmoins un certain nombres de difficultés, en particulier le grand nombre de variantes possibles qui ne diffèrent souvent que de quelques mots [Ogata and Goto, 2005]. Nous avons utilisé une représentation basée sur des réseaux de confusions, qui sont plus compacts que les graphes de mots, et plus *lisibles* pour l'humain. Avec une telle représentation, les actions correctives sont réduites à de simples actions d'édition du réseau : sélection d'un mot, suppression ou ajout d'une alternative manquante. Chaque action corrective effectuée sur le réseau de confusion est suivie d'un nouveau décodage, guidé par l'historique des corrections de l'utilisateur. L'objectif de ce re-décodage contraint par les corrections est d'améliorer la transcription et éventuellement de propager d'autres corrections dans son voisinage.

Le principe général du décodage contraint est d'extraire des réseaux de confusion partiellement édités un motif de phrase dans lequel les corrections apparaissent comme des mots. Les zones à re-décoder deviennent des jokers que le système doit trouver.

Techniquement, le décodage contraint est implémenté par l'algorithme de décodage guidé que j'avais proposé pour l'alignement de transcrip-

tion imparfaite [Lecouteux et al., 2006a] ou la combinaison de systèmes [Lecouteux et al., 2008]. Le schéma global du système de post-édition assistée est présenté dans la figure 3.3.

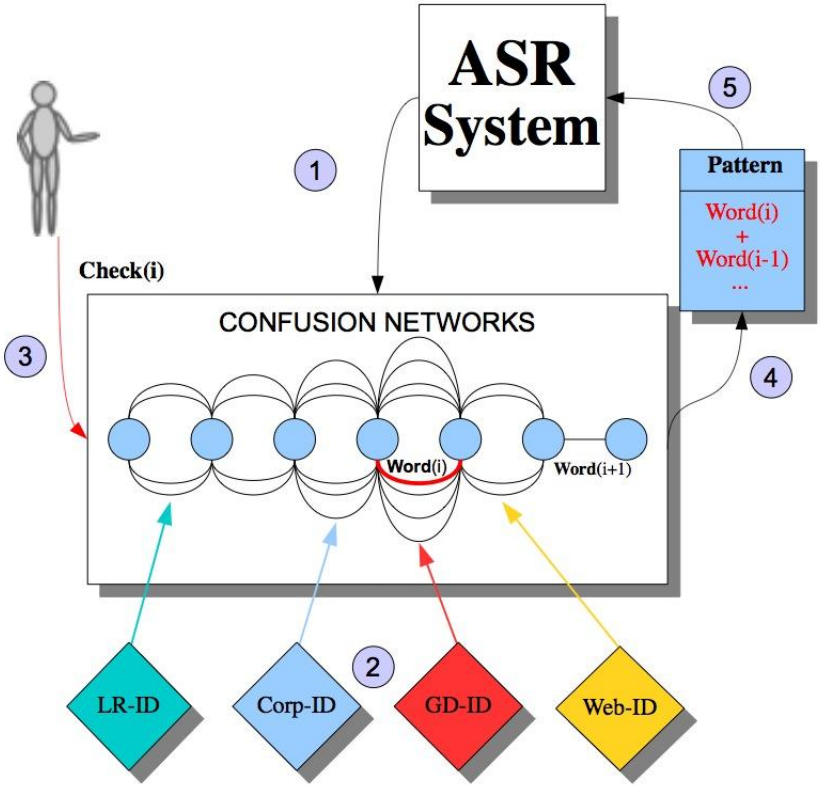


FIGURE 3.3 – Système de décodage interactif du SRAP [Senay et al., 2010a, Senay et al., 2010b] : l'utilisateur effectue des corrections sur le réseau de confusion ; ensuite un nouveau décodage est guidé par les nouvelles hypothèses.

Nous avons donc proposé un décodage guidé par le motif issu des corrections appliquées par le correcteur.

Considérant le scénario de correction incrémentale des transcriptions semi-automatiques, les zones dans lesquelles les corrections sont appliquées sont susceptibles de modifier sensiblement l'efficacité du décodage. Nous avons évalué différentes méthodes visant à guider le correcteur vers les zones les plus favorables au système en terme de réduction du taux d'erreurs mots (WER *Word Error Rate*) pour chaque acte correctif :

- Une correction guidée par la densité du graphe de mots : la largeur du graphe à un instant t est un indicateur pertinent de l'estimation des scores de confiances [Kemp and Schaaf, 1997], une "explosion" de la

largeur étant caractéristique d'une situation d'incertitude de l'algorithme. Le correcteur est orienté prioritairement vers les zones les plus explorées du graphe de mots.

- Une correction basée sur une consistance sémantique : le principe général est de considérer qu'un segment est inconsistant s'il est significativement distant de la dépêche du corpus la plus similaire (nous travaillions ici sur des *news*). La similarité du segment de transcription et des dépêches repose sur une mesure *Cosine* [Salton, 1989]. Cette consistance sémantique est détaillée dans la section 4.1.2.

- La consistance sémantique à l'aide du Web : l'idée de cette approche est de détecter des ruptures sémantiques dans une phrase en estimant à quel point les mots porteurs de sens sont cohérents entre eux et dans leur contexte.

  Pour mesurer à quel point un mot porteur de sens est cohérent avec son contexte, nous proposons d'utiliser la probabilité d'occurrence d'un mot dans un document Web, sachant que les mots de son contexte y apparaissent. Cette probabilité est formalisée dans l'équation 3.1, pour un mot w_i et une probabilité d'ordre n, et donc avec un contexte gauche de taille $n-1$, noté $\psi_i^n = w_{i-n-1}, \ldots, w_{i-1}$:

$$P_s(w_i|\psi_i^n) = \frac{WC(\psi_i^n, w_i)}{WC(\psi_i^n)} \tag{3.1}$$

  Avec $WC(\psi_i^n, w_i)$ le nombre de documents dans lesquels les mots ψ_i^n et w_i apparaissent quelle que soit leur position dans le document. Comme avec le modèle n-gramme classique, lorsqu'un mot n'apparaît dans aucun document web où figurent les mots de son contexte, on se replie sur la probabilité obtenue avec un contexte plus succinct, par un coefficient de repli déterminé empiriquement :

$$\hat{P}_s(w_i|\psi_i^n) = \begin{cases} P_s(w_i|\psi_i^n), \text{ si } WC(\psi_i^n, w_i) > 0 \\ \alpha \times P_s(w_i|\psi_i^{n-1}), \text{ sinon} \end{cases} \tag{3.2}$$

  La mesure de cohésion sémantique d'ordre n de la phrase est donc :

$$CS(w_1 \ldots w_i) = P_s(w_2|w_1) \times P_s(w_3|w_1, w_2) \times P_s(w_i|\psi_i^n) \tag{3.3}$$

Résultats

Les expériences ont été menées sur 8h extraites du corpus *dev* d'ESTER. Les détails sur le SRAP utilisé pour le décodage et les résultats plus détaillés sont donnés dans [Senay et al., 2010a, Senay et al., 2010b]. Les résultats présentés dans les tables 3.1 et 3.2 montrent que le décodage interactif permet une amélioration de l'efficacité de la correction, comparé à une approche manuelle. Le guidage sémantique est peu efficace dans les parties du corpus où le taux d'erreurs mots est faible, mais il devient plus efficient lorsque la transcription est de mauvaise qualité. Les corrections sont effectuées suivant différentes stratégies de guidage de la correction : Gauche-Droite (GD-ID), densité de graphe (DG-ID), correction sémantique basée sur des corpus (Corp-ID) ou basée sur le Web (Web-ID).

# c	1	3	10	20
Manuelle	25.22	22.98	17.23	9.44
GD-ID	24.28	**20.82**	**11.88**	**5.26**
DG-ID	26.58	25.38	16.62	11.76
Corp-ID	**23.90**	21.15	13.93	8.51
Web-ID	24.33	21.10	12.21	7.40

TABLE 3.1 – *WER selon le nombre d'actions correctives pour les segments dotés d'une transcription initiale à moins de 40% de WER.*

# c	1	3	10	20
Manuelle	55.91	54.05	47.81	40.14
GD-ID	54.95	49.77	37.71	**25.36**
DG-ID	57.51	53.52	44.05	36.99
Corp-ID	54.19	49.37	39.06	29.54
Web-ID	**51.88**	**48.32**	**37.49**	29.49

TABLE 3.2 – *WER selon le nombre d'actions correctives, pour les segments dotés d'une transcription initiales à plus de 40% de WER.*

3.3 Corrections automatiques basées sur les mesures de confiance

Généralement, les mesures de confiance sont utilisées pour l'adaptation des modèles acoustiques et l'apprentissage non supervisé, mais rarement pour

l'adaptation des modèles de langage. Certains travaux proposent d'utiliser les mesures de confiance pour modifier dynamiquement la pondération entre le modèle acoustique et le modèle linguistique pendant le processus de recherche, en fonction de la confiance de l'historique du modèle linguistique actuel. Dans [Fetter et al., 1996, Wessel et al., 1998], les auteurs utilisent la confiance des mots ou les probabilités *a posteriori* directement dans l'exploration du graphe, ce qui permet d'améliorer les performances de leur système.

3.3.1 Modèle cache sémantique pour l'adaptation linguistique

Nous abordons ici mes travaux présentés dans [Lecouteux et al., 2010] qui ont pour objectif d'introduire des mesures de confiance au sein du décodage d'un système de reconnaissance. L'objectif est d'exploiter tout le potentiel d'un premier décodage dans le processus de recherche suivant et de fournir une adaptation non supervisée du modèle de langue. Un modèle cache est appliqué pendant le processus de décodage, en exploitant à la fois des informations sémantiques calculées par une analyse de type *Latent Semantic Analysis* (LSA) et les mesures de confiance que nous avons présentées précédemment.

La stratégie utilisée consiste à se concentrer uniquement sur les mots mal reconnus, afin de réduire le bruit introduit. Dans un précédent article [Lecouteux et al., 2008], nous avions proposé un algorithme qui consistait à intégrer la sortie d'un SRAP auxiliaire dans l'algorithme de recherche d'un système primaire. Les modèles caches ont été introduits par [Kuhn and De Mori, 1990] : ils augmentent la probabilité des mots apparus récemment. L'hypothèse retenue étant que si un mot donné est utilisé, ce dernier a de fortes chances de réapparaître.

Nous en avons proposé une extension, consacrée au processus de décodage lui-même, en utilisant la sortie de la première passe pour piloter la seconde passe en fonction des scores de confiance des mots et des informations sémantiques associées.

Le module d'analyse sémantique

L'analyse sémantique latente (LSA) [Bellegarda, 2000] est une technique permettant d'associer des mots qui ont une corrélation sémantique au travers de plusieurs documents.

Dans notre SRAP, une séquence de mots sémantiquement cohérente peut être considérée comme peu probable par le modèle de langue en raison de

ses limites intrinsèques. Un score acoustique faible peut également éloigner une hypothèse correcte. Pour cette raison, nous intégrons un estimateur de consistance sémantique permettant de valider ou rejeter certaines hypothèses.

Dans nos expériences, le module LSA a été entraîné sur les données d'apprentissage du modèle de langage. Pour une meilleure couverture, le corpus a été lemmatisé et le vocabulaire réduit au lexique lemmatisé (environ 33K mots). De plus, une stop-liste a été appliquée pour filtrer les mots non porteurs de sens.

Lorsqu'un mot est présenté au module, ce dernier retourne les 100 meilleurs mots associés avec leurs scores de confiance LSA.

Décodage guidé avec LSA

L'utilisation du décodage guidé est indispensable dans ce contexte, car nos premières expériences ont montré qu'un modèle cache LSA seul générerait trop de bruit. Ainsi, le système est dirigé par ses hypothèses précédentes pour limiter les déviations des mots corrects. Le *trigger* LSA est appliqué uniquement sur les mots à faible confiance (< 0.5) : les mots corrects sont ainsi préservés. De plus, les mots associés à de faibles mesures de confiance sont sous-évalués, permettant au SRAP d'explorer des chemins alternatifs.

Le système final tel que détaillé dans la figure 3.4 fonctionne comme un modèle cache amélioré. Le DDA-LSA devient :

$$\phi(w_i) \geq 0.5 : \begin{cases} L(w_i|w_{i-2}..) = P(w_i|w_{i-2}..)^{1-\beta} \times \phi(w_i)^{\beta} \\ \beta = 0 \text{ si } \phi(w_i) \text{ est trouvé dans } H_{aux} \end{cases} \tag{3.4}$$

$$\phi(cw_i) < 0.5 : \begin{cases} L(w_i|w_{i-2}..) = P(w_i|w_{i-2}..)^{1-\alpha} \times \theta(w_i)^{\alpha} \\ \alpha = 0 \text{ si } \theta(w_i) \text{ non trouvé} \end{cases} \tag{3.5}$$

Où $L(w_i|w_{i-2}, w_{i-1})$ est le score linguistique résultant, $P(w_i|w_{i-2}, w_{i-1})$ est la probabilité du tri-gramme issu du modèle de langue, α et β sont des facteurs d'échelle déterminés empiriquement, $\phi(w_i)$ correspond au score de confiance du mot w_i, cw_i est le mot aligné après l'historique (w_{i-2}, w_{i-1}) dans la transcription auxiliaire et $\theta(w_i)$ est le score LSA de w_i.

Résultats

Nos expériences combinent une adaptation acoustique de type *Maximum Likelihood Linear Regression* (MLLR) pour la seconde passe (3x*Real Time* (RT)) avec le décodage guidé LSA, afin de tester la complémentarité du DDA-

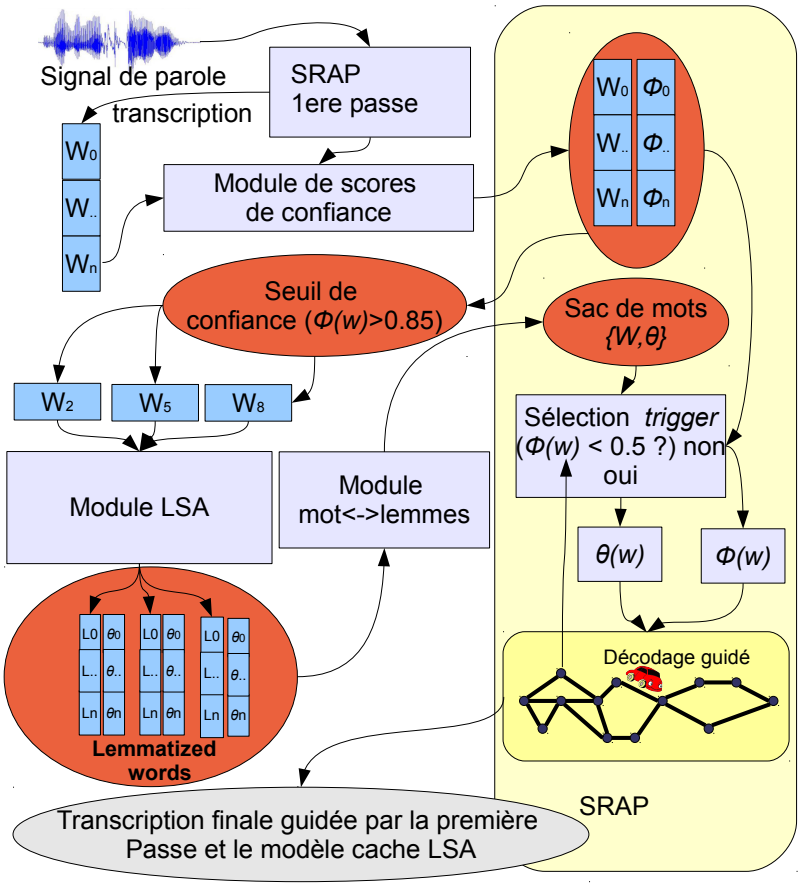


FIGURE 3.4 – Principe du décodage guidé par LSA. Figure extraite de [Lecouteux et al., 2010].

LSA avec l'adaptation acoustique. Les résultats présentés dans la table 3.3 montrent un WER réduit de 1.9% relatif. Ceci montre la complémentarité avec le processus d'adaptation acoustique. Nous observons le meilleur gain sur l'heure ayant les plus mauvaises performances en terme de décodage (4.6% relatifs).

Les expériences démontrent que cet algorithme permet d'améliorer le système initial et qu'il est complémentaire à l'adaptation des modèles acoustiques. L'enrichissement avec les scores de confiance oriente correctement l'algorithme de recherche et l'information sémantique sélectionne des chemins alternatifs corrects si les scores de confiance sont bas. Cette stratégie correspond à une adaptation non-supervisée et dynamique du modèle de langage.

Heure	P2 3RT	P2-LSA DDA 3RT
Classique 1h	20.8 %	20,5 %
Culture 1h	31.9 %	31.8 %
INTER 1h	22.0 %	21.6 %
INFO 2h	24.6 %	24.5 %
RFI 1h	26.0 %	25.5 %
RTM 2h	32.3 %	30.8 %

TABLE 3.3 – *baseline* en seconde passe 3xRT (P2 3xRT), décodage guidé par la sémantique avec adaptation acoustique en 3xRT (P2-LSA DDA 3xRT) [Lecouteux et al., 2010].

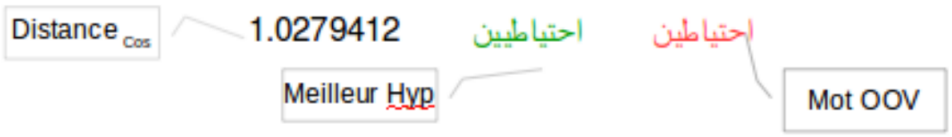
3.3.2 Détection de mots hors-vocabulaire et OCR

Dans le cadre du projet TRIDAN visant à travailler sur la traduction automatique de documents *océrisés*, nous avons été amenés à nous pencher sur la problématique des mots hors-vocabulaire en entrée d'un système de traduction automatique (TA). Dans cette situation, les mots hors vocabulaire sont liés à des erreurs de reconnaissance des caractères ou à une couverture insuffisante du corpus d'apprentissage. Ces mots sont importants à corriger car ils induisent une cascade d'erreurs entre les systèmes. Le système OCR utilisé est extrêmement proche d'un système de reconnaissance automatique de la parole : le système *Kaldi* [Povey et al., 2011b] a été détourné pour réaliser cette tâche. L'extraction de paramètres acoustiques a été remplacée par une extraction de paramètres issus de l'image. Le problème de détection de MHV devient donc un problème similaire à celui présenté dans la section 3.1. Ensuite, cette sortie est présentée en entrée d'un système de traduction probabiliste de type *phrase-based* construit à partir de *Moses* [Koehn et al., 2007a]. Le système OCR produit alors une liste des N-meilleures hypothèses combinées en un graphe qui sera décodé par *Moses*. Nous avons proposé deux approches [Bouzidi et al., 2017] :

Première approche : analyse de surface et sémantique

L'approche consiste à chercher et remplacer les MHV par des mots proches et connus de notre système de traduction avec deux critères de remplacement :

- Les mots qui se substitueront au mot hors-vocabulaire doivent posséder un contexte d'utilisation proche de celui du mot inconnu (distance *cosine* entre représentations vectorielles des mots).



- Les mots qui remplaceront le mot hors-vocabulaire doivent avoir une forme de surface, se basant sur une distance de Levenshtein, relativement proche (figure 3.5).

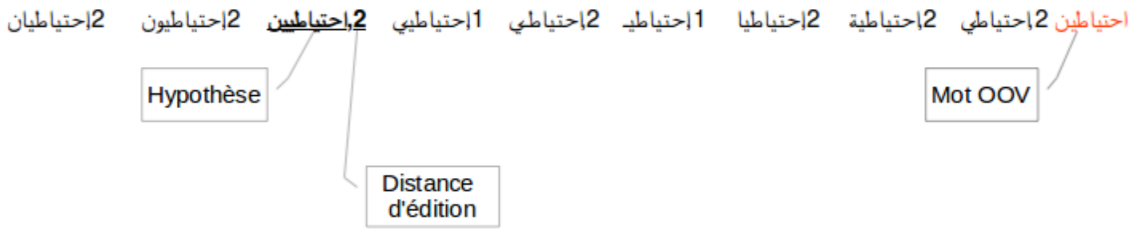


FIGURE 3.5 – Le mot hors-vocabulaire doit avoir une forme de surface, se basant sur une distance de Levenshtein, relativement proche. Figure extraite de [Bouzidi et al., 2017].

Seconde approche : augmentation du graphe de mots

Dans cette approche nous réutilisons la même mesure de similarité composite (forme de surface et contexte). Mais le fait d'avoir un graphe permet d'ajouter plusieurs mots candidats en remplacement d'un MHV. Ainsi, les M mots les plus similaires en fonction de la distance *cosine* relative au MHV sont sélectionnés.

La figure 3.6 montre un exemple où le mot inconnu est substitué par 4 nouveaux arcs qui représentent les meilleurs candidats au remplacement du mot inconnu. Dans ce cas, le MHV est conservé dans le graphe et le décodeur choisira quel chemin sera le plus prometteur.

Résultats

Corpus d'expérimentation

Les corpus de développement et de test sont des extraits de journaux issus de la base MAURDOR : ce sont des images numérisées de journaux en arabe pour lesquelles on dispose de la transcription exacte et automatique (15%

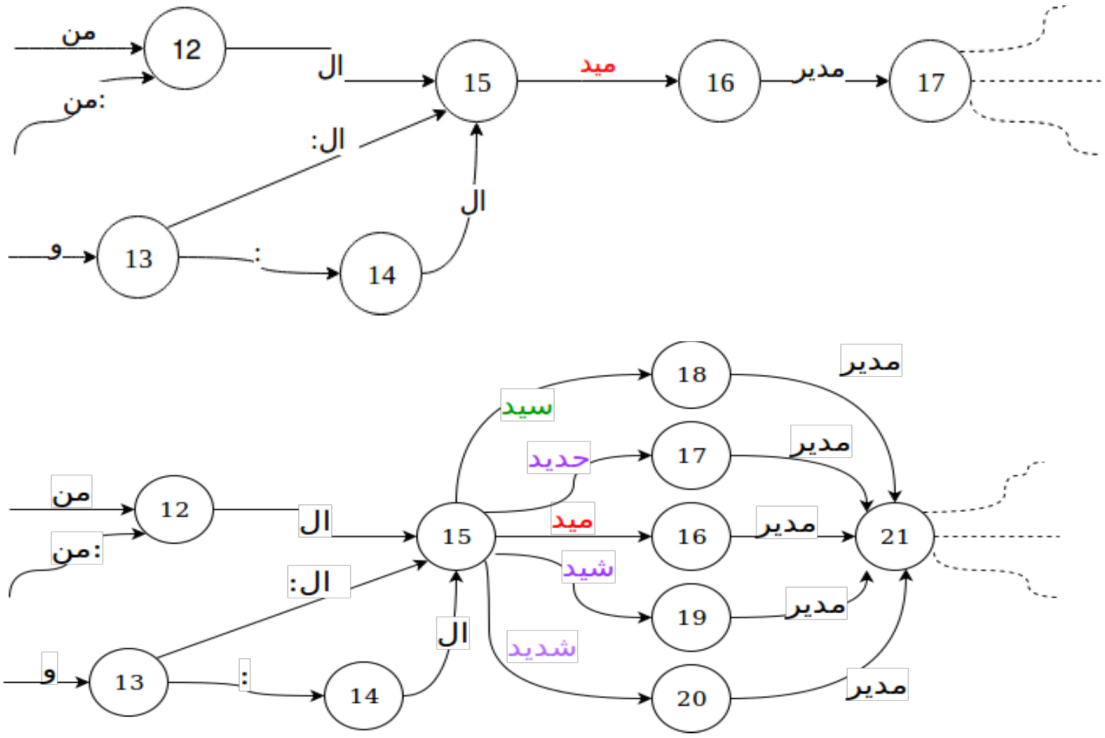


FIGURE 3.6 – Augmentation du graphe avec des mots proches du MHV (en rouge).Figures extraites de [Bouzidi et al., 2017].

d'erreur du système OCR). Au final nous avons 250 phrases pour le dev et 267 pour le test.

Expériences

Système	Dev (BLEU)	Test (BLEU)	Dev (%MHV)	Test (%MHV)
Baseline	21,17	16,41	1,96	2
+traitement MHV	21,34	16,59	1,04	1,04
+Graphe	24,63	21,75	1,43	1,62
+traitement MHV+Graphe	**24,9**	**21,91**	**0,2**	**0,2**

TABLE 3.4 – Résultats avec détection et correction des MHV extraits de [Bouzidi et al., 2017].

Les expériences dont les résultats sont présentés dans la table 3.4 montrent une approche qui améliore les résultats de traduction et réduit drastiquement le nombre de mots hors vocabulaire. Nous observons que l'augmentation des graphes d'hypothèses permet les meilleurs gains. Peu de travaux antérieurs

avaient démontré l'efficacité d'un couplage plus étroit au niveau de la cascade OCR/TA.

3.4 Conclusion

Dans ce chapitre nous avons mis en évidence de quelles manières diverses mesures de confiance pouvaient être exploitées pour améliorer des systèmes de reconnaissance automatique de la parole ou de reconnaissance automatique de caractères. Nous avons exploré à la fois l'interaction avec l'utilisateur où les mesures de confiance l'aident dans ses choix de post-édition et l'exploitation automatique des mesures de confiance. Actuellement, la confiance sur les sorties d'un système (que ce soit en parole, traduction ou autre) est toujours une question ouverte. Même avec les progrès liés aux réseaux de neurones, les systèmes n'ont aucun "recul" sur les informations qu'ils produisent. Par exemple, un système de traduction excellent peut produire des traductions totalement incohérentes sur des phrases particulières ou plus dramatiquement un système de conduite automatique confondra un camion bleu avec le ciel [1]. Les techniques ont par ailleurs évolué avec les réseaux de neurones profonds où les paramètres à prendre en considération pour évaluer la confiance diffèrent totalement des anciens systèmes. Dans le chapitre suivant, nous abordons ces questions sous un autre angle : la prédiction de performance en fonction des données.

1. https ://www.lemonde.fr/economie/article/2021/08/16/le-systeme-d-autopilotage-des-tesla-vise-par-une-enquete-apres-une-serie-de-onze-accidents-aux-etats-unis_6091586_3234.html

Chapitre 4

Prédiction d'indexabilité et de performance

Encadrement : Zied Elloumi

Sélection de publications relatives à ce chapitre :

Z. Elloumi, **B. Lecouteux**, O. Galibert, L. Besacier. Prédiction de performance des systèmes de reconnaissance automatique de la parole à l'aide de réseaux de neurones convolutifs. TALN 2018.

Z. Elloumi, L. Besacier, O. Galibert, J. Kahn, **B. Lecouteux**. ASR performance prediction on unseen broadcast programs using convolutional neurol networks. ICASSP 2018.

Z. Elloumi, L. Besacier, O. Galibert, **B. Lecouteux**. Analyzing Learned Representations of a Deep ASR Performance Prediction Model. Blackbox NLP Workshop at EMNLP 2018.

G. Senay, G. Linarès, **B. Lecouteux** (2011) « A segment-level confidence measure for spoken document retrieval. ICASSP 2011.

Ce chapitre aborde des travaux ayant des objectifs similaires mais pour des tâches différentes. Dans un premier temps nous aborderons la prédiction de l'indexabilité qui vise à déterminer dans quelle mesure un document qui a été automatiquement transcrit pourra être retrouvé en fonction des erreurs potentielles qu'il contiendra. Ces travaux [Senay et al., 2010b] ont été réalisés en collaboration avec Grégory Senay et Georges Linarès : mes principales contributions ont porté sur les mesures de confiance. Dans la seconde partie nous aborderons les travaux réalisés dans le cadre de la thèse de Zied Elloumi [Elloumi, 2019] que j'ai co-encadrée avec Laurent Besacier. Ces travaux [Elloumi et al., 2018a, Elloumi et al., 2018b, Elloumi et al., 2018c] ont porté sur la prédiction de performance dans le cadre de la reconnaissance automatique de la parole. La tâche consiste à estimer, en s'appuyant sur le signal, quels résultats pourra atteindre un système de reconnaissance de la parole (aussi bien en terme de WER que de coûts de développement, quantités de données nécessaires à l'apprentissage etc.).

4.1 SRAP et prédiction de l'indexabilité

Il est possible d'associer la reconnaissance automatique de la parole à des systèmes de recherche d'information (RI). Cette option nécessite de surmonter la principale difficulté liée aux erreurs de reconnaissance des systèmes de RI. Ceux-ci souffrent d'un manque de robustesse préjudiciable. Dans des cas inattendus les taux d'erreurs mots peuvent s'avérer supérieurs à 30 %, et détériorer significativement la précision de la recherche vocale [Whittaker et al., 2002, Oard et al., 2004, Hansen et al., 2005]. Toutefois dans des conditions contrôlées où les SRAP ont un taux d'erreur d'environ 10% WER, la campagne TREC-SDR a montré que ces erreurs n'ont qu'un léger impact sur les moteurs de RI [Garofolo et al., 2000].

De nombreuses études se sont concentrées sur des méthodes d'indexation tolérantes aux erreurs, basées sur des représentations de mots ou des N-meilleures hypothèses de sorties de SRAP [Siegler, 1999, Saraclar, 2004, Kurimo and Turunen, 2005, Chelba et al., 2007, Chang et al., 2008] et le traitement des mots hors-vocabulaire.

Pour les applications industrielles, il est possible d'identifier de manière réaliste les segments de parole où le système RAP échoue, non seulement en termes de WER mais aussi en considérant l'objectif final. Les segments erronés peuvent alors être vérifiés et corrigés par un opérateur humain. Dans ce scénario semi-automatique, la disponibilité d'un outil d'auto-diagnostic aidant

l'opérateur à identifier les segments corrompus est critique pour le coût global du processus d'indexation.

Nous avons proposé une méthode qui permet de prédire l'impact d'une transcription erronée sur la performance globale d'un SRAP.

Étant donné que les erreurs au sein d'un segment peuvent avoir un impact sur tous les résultats de recherche, l'estimation de l'indexabilité $Idx(s)$ d'un segment s est calculée par un processus en 3 étapes :

1. le segment vocal ciblé s est automatiquement transcrit par le SRAP
2. pour chaque requête du test, la recherche est effectuée sur l'ensemble de la base de données en utilisant des transcriptions correctes pour tous les segments, à l'exception de s qui est automatiquement transcrit
3. les rangs résultants sont comparés à ceux obtenus en recherchant l'ensemble complet de transcriptions de référence. L'indexabilité $Idx(s)$ du segment s est obtenue en calculant la F-mesure sur les 20 premiers segments classés, par rapport au classement de référence.

Cet algorithme permet d'évaluer l'impact individuel de la transcription du segment ciblé sur le processus global. La méthode proposée vise à prédire l'impact des erreurs de reconnaissance sur le processus d'indexation. Pour ce faire, elle combine des mesures de confiance au niveau des mots et un indice de compacité sémantique sur la meilleure hypothèse issue du SRAP. Une combinaison des deux mesures est ensuite réalisée en utilisant un perceptron multicouche.

4.1.1 Mesure de confiance du SRAP

Les scores de confiance sont calculés en deux étapes. La première consiste à extraire des caractéristiques de bas niveau liées à la topologie du graphe de recherche et à l'acoustique, ainsi que des caractéristiques de haut niveau liées à la linguistique. Dans la deuxième étape, une première hypothèse de détection d'erreur est produite par un classificateur basé sur l'algorithme de boosting. La méthode employée est similaire à celle présentée dans la section 3.1. La principale différence étant que nous utilisons l'algorithme de classification par boosting afin de combiner les caractéristiques des mots pour estimer la probabilité qu'un mot soit correct ou non.

4.1.2 Indice de Compacité Sémantique

L'utilisation d'informations de niveau sémantique pour la prédiction de l'indexabilité est motivée par le fait qu'une requête cible généralement les

documents en fonction de leur contenu sémantique. Certains travaux ont proposé d'utiliser ces caractéristiques de haut niveau pour l'estimation des mesures de confiance [Cox and Dasmahapatra, 2002, Hakkani-Tür et al., 2005].

Notre proposition est d'estimer un indice de compacité sémantique $SCI(s)$ pour chaque segment s et de l'utiliser comme caractéristique d'entrée du prédicteur. Ce score de segment est obtenu en faisant la moyenne des corrélations sémantiques locales $sc(w_i, w_j)$ au niveau des paires de mots (w_i, w_j) estimées sur un large corpus.

Nous nous concentrons sur les corrélations à court terme entre les mots significatifs. Les termes restants sont lemmatisés à la fois dans le corpus et dans les segments de transcription. Ensuite, les scores sémantiques des paires de mots sont calculés en utilisant les fréquences de co-occurrences de lemmes pondérées par un indice $TF - IDF$:

$$sc(w_i,w_j) = \frac{TF(l_i,c).IDF(l_i).\delta^c(w_j)+TF(l_i,c).IDF(l_j).\delta^c(w_i)}{} \tag{4.1}$$

où l_i est la forme lemmatisée du mot w_i , $TF(l_i, c)$ la fréquence du lemme l_i dans le contexte c, $IDF(l)$ la fréquence inverse du lemme l sur l'ensemble du corpus, $\delta^c(w) = 1$ si $w \in c$ et $\delta^c(w) = 0$ si $w \notin c$.

La compacité sémantique $sci(c)$ est estimée sur une fenêtre glissante de 5 lemmes et à chacun correspond un contexte c :

$$sci(c) = \sum_{c_k \in s} \sum_{(w_i,w_j)\in c_k} \sqrt{sc(w_i, w_j).\frac{IDF(w_i)IDF(w_j)}{\sum_{k=1}^{n} IDF(w_k)}} \tag{4.2}$$

Dans nos expériences, les statistiques sont calculées sur la partie française du corpus de Wikipédia qui couvre un grand nombre de sujets et de thèmes.

Résultats

Les taux d'erreur de prédiction (*Prediction Error Rate* PER) ont été évalués sur un corpus de test de 7 heures issu d'ESTER 2 [Galliano et al., 2006]. Afin d'estimer la contribution individuelle de chaque caractéristique, nous avons entraîné le classifieur sur : la mesure de confiance (CM), la compacité sémantique (SCI) et la combinaison de $CM + SCI$.

Les distorsions entre la prédiction de l'indexabilité $PIdx$ et l'indexabilité effective Idx sont évaluées avec les métriques suivantes :

$$Q1 = \frac{1}{\tau}\sum_{j=1}^{\tau}\frac{|PIdx(j) - Idx(j)|}{Idx(j)} \tag{4.3}$$

$$RMS = \sqrt{\frac{1}{\tau}\sum_{j=1}^{\tau}(PIdx(j) - Idx(j))^2} \tag{4.4}$$

$Q1$ et RMS présentent respectivement la distorsion générale et la déviation standard. τ est le nombre de segments. Les résultats présentés dans la table 4.1 montrent la complémentarité des deux indices, la combinaison surpassant d'environ 13% les PER obtenus par le meilleur prédicteur utilisé seul.

	CM	SCI	$CM + SCI$
PER : $Q1$	1,65	1,74	1,59
PER : RMS	0,25	0,32	0,22

TABLE 4.1 – Erreur moyenne de prédiction avec des métriques de distorsion ($Q1$) et déviation standard (RMS), en utilisant respectivement la mesure de confiance seule (CM), l'indice de compacité sémantique SCI seul et la combinaison des 2 indices ($CM + SCI$).

4.2 SRAP et prédiction de performance

Prédire la performance de la reconnaissance automatique de la parole pour de nouveaux enregistrements vocaux représente un Graal dans le domaine de la RAP. Une telle tâche aide à comprendre les facteurs liés à la performance d'un système. D'un point de vue technique, la prédiction de performance est utile dans des cadres applicatifs où des systèmes de transcription doivent être rapidement mis en place (ou adaptés) pour de nouveaux types de documents (prédiction des courbes d'apprentissage, estimation de la quantité de données d'adaptation nécessaires pour atteindre une performance acceptable, etc.). La prédiction de performance à partir de documents non vus diffère de l'estimation de la confiance : les systèmes de mesure de confiance permettent de détecter les parties correctes ainsi que les erreurs dans une sortie de SRAP et sont généralement adaptés à un système particulier et à des types de documents connus. La prédiction de performance se concentre sur des documents non vus et le diagnostic est fourni à une granularité plus large, au niveau du document par exemple. Dans le cadre de la prédiction de performance, nous considérons que nous n'avons pas accès au SRAP étudié (pas de graphe, pas N-meilleures

hypothèses, aucun élément interne du SRAP) et il sera considéré comme une boîte noire. Nous partons de l'hypothèse que les systèmes de prédiction de performance ne doivent utiliser que les transcriptions et le signal comme entrée afin d'évaluer la qualité de transcription correspondante. Les transcriptions de référence (humaines) ne sont nécessaires que pour l'entraînement du système de prédiction.

La thèse de Zied Elloumi que j'ai co-encadrée avec Laurent Besacier a notamment porté sur la prédiction de performance de SRAP à partir d'enregistrements qui n'avaient encore jamais été observés ; Cette section présente une partie de ces travaux [Elloumi et al., 2018a, Elloumi et al., 2018c, Elloumi, 2019].

4.2.1 Corpus dédié à la tâche de prédiction de performance

Dans le cadre de la thèse de Zied Elloumi, nous avons rassemblé un corpus français large et hétérogène (contenant de la parole spontanée et non spontanée) dédié à cette tâche et avons proposé un protocole d'évaluation.

Le corpus $Train_{pred}$ contient 75000 paires $\{sortie_{SRAP}, Performance\}$, le corpus $Test_{pred}$ est composé de 6800 sorties de SRAP pour lesquelles nous essayons de prédire la performance de transcription associée. Les transcriptions de référence (humaines) pour le $Test_{pred}$ sont uniquement utilisées pour évaluer la qualité de la prédiction.

Les données utilisées proviennent de différentes collections d'émissions françaises :

- Des données issues de REPERE : 54 heures de transcription d'émissions (spontanées, telles que des débats) et des journaux télévisés.
- Les données d'ESTER (1 et 2) contenant 111h de transcription audio, provenant principalement de programmes français et africains (mélange de discours préparés et de discours plus spontanés : discours de présentateur, interviews, reportages).
- Les données du projet ETAPE qui comprennent 37h de programmes de radio et de télévision (principalement spontanée avec des locuteurs dont les dialogues se chevauchent).
- Des données issues de Quaero contenant 41h de discours diffusés par différentes radios et télévisions françaises sur des sujets variés.

Les données contiennent ainsi des discours non-spontanés (NS) et spontanés (S). Les données utilisées pour entraîner notre système de reconnaissance de la

parole sont sélectionnées parmi les différents styles de discours non-spontanés qui correspondent principalement aux journaux télévisés. Les données utilisées pour la prédiction des performances sont un mélange des deux styles de discours (S et NS).

4.2.2 Système *baseline*

Afin d'obtenir un système de prédiction de référence, nous avons adapté le système TranscRater [Jalalvand et al., 2016] pour le français. Il utilise des traits explicites permettant de prédire la performance de chaque entrée et se base sur l'algorithme *Extremely Randomized Trees* [Geurts et al., 2006] pour l'apprentissage. Les traits sont sélectionnés avec l'algorithme *Randomized Lasso* [Meinshausen and Bühlmann, 2010]. Les paramètres du modèle sont optimisés via une recherche exhaustive qui minimise l'erreur absolue moyenne (*Mean Absolute Error*, MAE) entre les WER prédits et leur référence.

TranscRater permet d'extraire des traits de quatre types :

- morphosyntaxiques (POS)
- provenant du modèle de langue (LM)
- lexicaux (LEX)
- acoustiques (SIG)

Ce système à base de régression a été utilisé comme *baseline* par rapport aux approches à base de réseaux de neurones que nous proposerons.

4.2.3 Prédiction de performance basée sur des réseaux de neurones convolutifs

La principale contribution des travaux de Zied Elloumi s'articule autour de la comparaison objective entre une prédiction de performance basée sur la régression (caractéristiques techniques) et une nouvelle stratégie basée sur les réseaux de neurones convolutifs (caractéristiques apprises). Afin de réaliser notre prédiction, nous nous sommes appuyés sur les deux modalités qui sont à notre disposition lorsque nous avons un corpus : son signal audio et sa transcription.

Pour représenter le texte, nous nous inspirons de l'architecture proposée par [Kim, 2014] . L'entrée textuelle correspond à un tour de parole qui a été complété à N mots et représenté via une matrice $EMBED$ de taille $N \times M$ (M = dimension du vecteur de la représentation continue d'un mot) de telle sorte que chaque ligne de la matrice $EMBED$ est associée à une représentation vectorielle d'un mot.

Pour l'entrée du signal, nous nous inspirons de l'architecture proposée dans [Dai et al., 2017]. C'est un CNN composé de 17 couches de convolution + *max-pooling* suivies d'une opération d'agrégation (*global average pooling*) puis de 3 couches cachées complètement connectées.

Afin de joindre les données textuelles et acoustiques, nous combinons les 2 dernières couches cachées *(CNN RAW-SIG)* et *(CNN EMBED)* en les concaténant pour obtenir une nouvelle couche cachée.

L'architecture globale du modèle proposé est présentée dans la figure 4.1.

Contrairement aux traits de l'approche par régression, les traits pour le CNN textuel sont extraits et entraînés via les représentations vectorielles des mots. Ils sont ensuite appris par le réseau neuronal jusqu'à convergence.

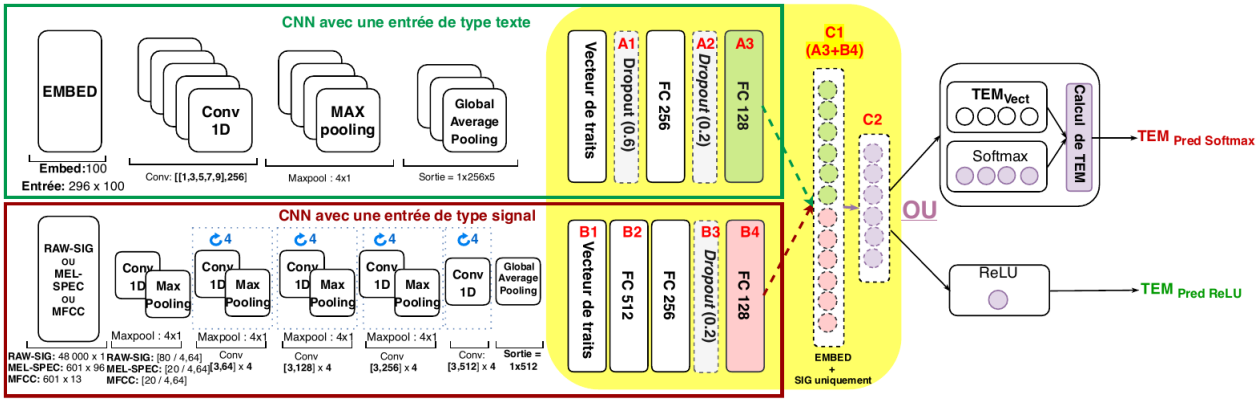


FIGURE 4.1 – Système de prédiction de performances proposé par Zied Elloumi [Elloumi et al., 2018a, Elloumi et al., 2018c]. Afin de joindre les données textuelles et acoustiques, nous combinons les 2 dernières couches cachées *(CNN RAW-SIG)* et *(CNN EMBED)* en les concaténant pour obtenir une nouvelle couche cachée.

Résultats

Les résultats des expériences menées sur le corpus décrit précédemment, sont présentés dans la table 4.2 et montrent que la prédiction via CNN est meilleure que le système *baseline* TranscRater aussi bien au niveau de la mesure de Kendall que du MAE. L'utilisation jointe en entrée du signal et du texte n'apporte pas de gain avec l'approche régressive, tandis qu'elle permet des améliorations substantielles en utilisant des CNN. Nous avons également montré que les CNN prédisent bien mieux la distribution des erreurs mots sur une

collection d'enregistrements, tandis que TranscRater prédit une distribution quasi-gaussienne autour du WER moyen observé dans l'apprentissage : ce constat est clairement illustré sur la figure 4.2.

Modèle	Input	MAE	Kendall
Caractéristiques textuelles (TXT)			
TranscRater	POS + LEX + LM	22,01	**44,16**
CNN$_{Softmax}$	EMBED	**21,48**	38,91
CNN$_{ReLU}$	EMBED	22,30	38,13
Caractéristiques acoustiques (SIG)			
TranscRater	SIG	25,86	23,36
CNN$_{Softmax}$	RAW - SIG	25,97	23,61
CNN$_{ReLU}$	RAW - SIG	26,90	21,26
CNN$_{Softmax}$	MEL - SPEC	29,11	19,76
CNN$_{ReLU}$	MEL - SPEC	26,07	24,29
CNN$_{Softmax}$	MFCC	**25,52**	**26,63**
CNN$_{ReLU}$	MFCC	26,17	25,41
Caractéristiques textuelles et acoustiques (TXT + SIG)			
TranscRater	POS + LEX + LM + SIG	21,99	45,82
CNN$_{Softmax}$	EMBED + RAW - SIG	**19,24**	**46,83**
CNN$_{ReLU}$	EMBED + RAW - SIG	20,56	45,01
CNN$_{Softmax}$	EMBED + MEL-SPEC	20,93	40,96
CNN$_{ReLU}$	EMBED + MEL-SPEC	20,93	44,38
CNN$_{Softmax}$	EMBED + MFCC	19,97	44,71
CNN$_{ReLU}$	EMBED + MFCC	20,32	45,52

TABLE 4.2 – Résultats de prédiction [Elloumi et al., 2018c] en fonction des différents types de traits utilisés et comparés à un système *baseline* à base de régression (TranscRater).

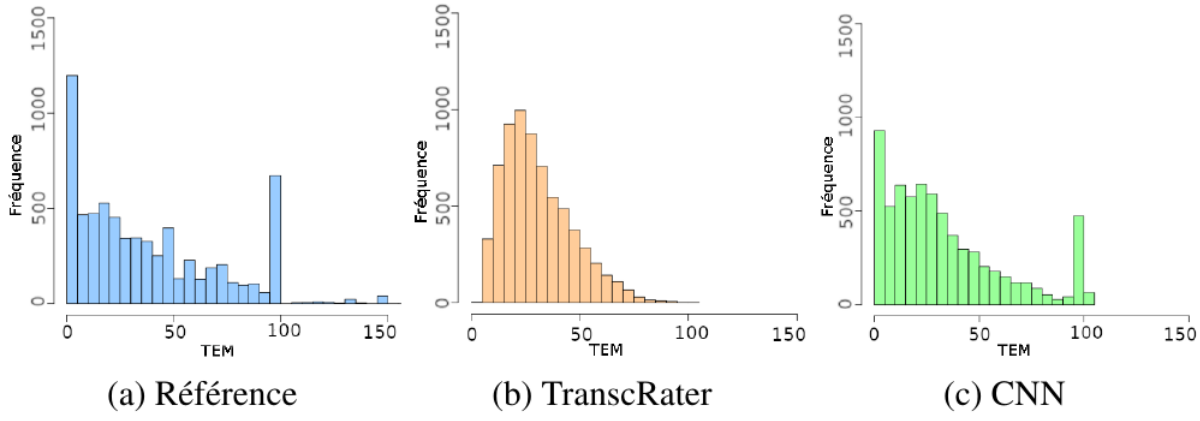


FIGURE 4.2 – Les CNN prédisent bien mieux la distribution des taux d'erreurs de mots sur une collection d'enregistrements [Elloumi et al., 2018c].

4.2.4 Représentations apprises par les réseaux de neurones lors de la prédiction de performance

Lors du doctorat de Zied Elloumi, nous nous sommes également intéressés aux représentations qui étaient capturées par les réseaux de neurones [Elloumi et al., 2018b, Elloumi et al., 2018c]. Cela permet de comprendre quelles particularités aident à prédire la qualité d'un système. Pour réaliser ces analyses, nous nous sommes inspirés des recherches de [Belinkov and Glass, 2017] : notre modèle pré-entraîné est utilisé pour générer des représentations au niveau du tour de parole. Ainsi, nous analysons les représentations présentes dans les couches supérieures de notre réseau. Ces représentations sont extraites puis utilisées pour entraîner un classifieur et résoudre des tâches de classification comme :

- la détection du type d'émission (journal, interview etc.)
- la détection d'accent (locuteurs natifs et non-natifs en utilisant les annotations des locuteurs fournies avec nos données afin d'étiqueter les données)
- la détection de style (spontané ou non)

Les performances des classifieurs annexes permettent de savoir quelles informations (style, accent, émission) sont le mieux capturées par telle ou telle couche. Afin d'avoir une analyse visuelle, nous avons projeté les représentations dans un espace à 2 dimensions via l'algorithme *t-SNE* [Maaten and Hinton, 2008].

La table 4.3 rassemble les résultats des classifieurs annexes sur les différentes couches. Ils montrent que le modèle capture des informations sur les trois modalités cherchées durant l'apprentissage du système.

Une représentation visuelle des informations capturées par la couche C2 (se référer à la figure 4.1 pour identifier les différentes couches) est présentée dans la figure 4.3 : la parole non spontanée est bleue tandis que la parole spontanée est rose. La couche C2 produit des *clusters* montrant la parole non spontanée dans la partie inférieure droite de l'espace 2D : la couche C2 véhicule donc une information relative au style de parole.

Suite à ces résultats, nous avons étudié le potentiel d'un apprentissage multi-tâche incluant ces trois informations lors de l'entraînement. Cela améliore la prédiction de WER et génère une prédiction correcte concernant l'accent du locuteur, le style de parole ou encore le type d'émission.

Couche	Dim.	ÉMISSION	STYLE	ACCENT
		EMBED		
A1	1 280	**57,12**||-	**80,72**||68,99	**70,75**||66,54
A2	256	54,89||-	80,01||**69,56**	69,30||69,43
A3	128	51,04||-	79,23||68,27	68,25||**70,89**
		RAW - SIG		
B1	512	**42,35**||-	**72,92**||**58,64**	64,60||**55,85**
B2	512	41,22||-	72,20||58,41	64,44||54,84
B3	256	41,22||-	72,38||58,44	64,50||54,65
B4	128	40,77||-	72,38||58,52	**64,74**||54,87
		EMBED + RAW - SIG		
C1 (A3+B4)	256	**57,04**||-	**81,29**||70,36	**71,41**||**65,98**
C2	128	53,06||-	79,62||**70,55**	70,01||65,20
Aléatoire	-	**20,00**	**50,00**	**50,00**

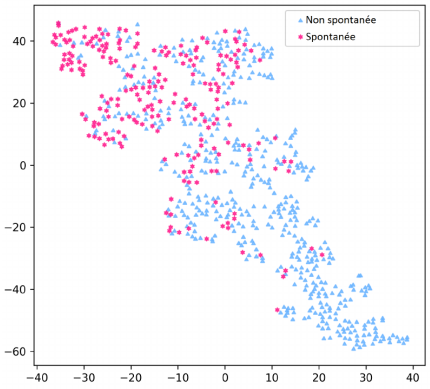


TABLE 4.3 – Analyse de l'information capturée par les différents classifieurs à partir de l'étude des performances des systèmes de classification style accent et émission en termes de taux de bonne classification et visualisation des informations capturées par la couche cachée C2. Figure et table extraites de [Elloumi et al., 2018c].

4.3 Conclusion

Dans ce chapitre nous avons présenté les travaux axés sur la prédiction d'indexabilité, puis de performance. Dans les deux cas nous avons établi qu'il était possible d'obtenir des indices intéressants représentatifs de la difficulté à traiter des données particulières. Pour ce type d'étude, l'une des grandes difficulté est la mise en place des protocoles et corpus qui permettent d'évaluer les méthodes de prédiction. Cet aspect n'a pas été abordé dans ce chapitre, mais une grande partie du travail dans la thèse de Zied Elloumi a porté sur les métriques d'évaluation pertinentes ainsi que sur la mise en place de protocoles ou corpus fiables et de systèmes à état-de-l'art. Par ailleurs, l'analyse des données capturées par les réseaux de neurones est extrêmement intéressante dans la perspective d'acquérir une meilleure compréhension de ce que perçoit (ou non) le système d'apprentissage. Le chapitre suivant revient sur les mesures de confiance, mais dédiées aux systèmes de traduction automatique.

Chapitre 5

Mesures de confiance pour la traduction automatique

Encadrements : NQ. Luong, Loïc Vial

Sélection de publications relatives à ce chapitre :

L. Vial, **B. Lecouteux**, D. Schwab, H. Le, L. Besacier. The LIG system for the English-Czech Text Translation Task of IWSLT 2019. IWSLT 2019.

NQ. Luong, L. Besacier, **B. Lecouteux**. Find The Errors, Get The Better : Enhancing Machine Translation via Word Confidence Estimation. Natural Language Engineering 2017

NQ. Luong, L. Besacier, **B. Lecouteux**. Towards Accurate Predictors of Word Quality for Machine Translation : Lessons Learned on French-English and English-Spanish Systems. KSE journal 2015.

NQ. Luong, L. Besacier, **B. Lecouteux**. Word Confidence Estimation for SMT N-best List Re-ranking. Workshop on Humans and Computer-assisted Translation (HaCaT) during EACL 2014.

Peu après ma prise de poste en 2011, Laurent Besacier m'a proposé de co-encadrer une thèse portant sur les mesures de confiance appliquées à la traduction automatique. Cette opportunité a été l'occasion de poursuivre mes travaux sur les mesures de confiance dans un nouveau domaine de recherche : la traduction. La première partie de ce chapitre porte donc sur des mesures de confiance spécifiques à la traduction automatique et une mise en œuvre de ces dernières pour mettre en place un décodage multi-passe de traduction comme cela se faisait couramment dans le cadre de la reconnaissance automatique de la parole. Ces travaux [Luong et al., 2013a, Luong et al., 2013b, Luong et al., 2014a, Luong et al., 2014c, Luong et al., 2015] ont été portés par Ngoc Quang Luong dans le courant de sa thèse [Luong, 2014] encadrée par Laurent Besacier et moi-même. La seconde partie de ce chapitre portera sur une partie des travaux de Loic Vial [Vial et al., 2019d, Vial, 2020]. Il a été encadré par Didier Schwab et moi-même. Sa thèse de doctorat a principalement porté sur la désambiguïsation lexicale et nous présenterons ses travaux permettant d'introduire la désambiguïsation lexicale au sein d'un système de traduction automatique, ce qui peut s'apparenter à une sorte de mesure de confiance.

5.1 Traduction automatique du texte

5.1.1 Mesures de confiance spécifiques à la TA

Lors de l'encadrement de Ngoc Quang Luong, nous avons travaillé sur les mesures de confiance intégrées au décodage de systèmes de traduction automatique. Les décodeurs peuvent en effet être améliorés si des informations supplémentaires sont introduites dans la phase de décodage. Les approches classiques se basent généralement sur de la post-édition, du réordonnancement d'hypothèses ou le redécodage avec d'autres modèles [Ueffing and Ney, 2007, Sokolov et al., 2012], etc. La post-édition est avant tout une tâche assistée par l'humain. En ce qui concerne le réordonnancement, d'autres traits sont intégrés aux scores des modèles existants pour sélectionner une meilleure hypothèse. Nous avons privilégié un processus de redécodage intervenant directement au niveau du graphe de recherche du décodeur. Dans la littérature, le décodage à deux passes a été exploité de différentes manières : on peut par exemple entraîner des modèles de langue et les combiner avec les scores des modèles existants pour ré-ordonner les hypothèses [Kirchhoff and Yang, 2005]. Il est également possible, dans une seconde passe [Och, 2003, Nguyen et al., 2011], d'utiliser des scores tant au niveau des phrases qu'au niveau des mots. Nous

nous sommes concentrés sur une seconde passe où le coût de chaque hypothèse présente dans le graphe de recherche sera mis à jour en fonction de la qualité prédite par un système d'estimation de mesures de confiance. Dans un décodage à une passe, le décodeur recherche parmi les chemins complets celui qui permet d'obtenir le coût optimal : le coût de l'hypothèse est un score composite, issu de divers modèles du système de traduction (ré-ordonnancement, traduction, LMs, etc.). Bien que les N-meilleures hypothèses devancent les autres hypothèses en termes de scores, il n'y a pas d'indice certain qu'elles seront les plus proches des références humaines. Nous avons introduit une seconde passe où des scores liés à la prédiction de la confiance des mots sont intégrés dans le graphe de recherche afin de pouvoir réévaluer la meilleure hypothèse.

Mesures de confiance basées sur un modèle CRF

La première étape consiste à appliquer des mesures de confiance sur les N-meilleures hypothèses pour leur attribuer des étiquettes de qualité ("Good" ou "Bad") ainsi que des probabilités de confiance au niveau du mot. Les mesures de confiance calculées ont évolué par rapport à celles présentées dans le chapitre précédent : nous abordons désormais le problème de mesure de confiance comme un processus d'étiquetage de séquences. Pour cela nous avons utilisé les champs aléatoires conditionnels (*Conditional Random Field* CRF) pour l'apprentissage de notre modèle, avec la boîte à outils WAPITI [Lavergne et al., 2010] :

Soit $X = (x_1, x_2, \ldots, x_N)$ la variable aléatoire sur la séquence de données à étiqueter, $Y = (y_1, y_2, \ldots, y_N)$ la séquence de sortie obtenue après la tâche d'étiquetage. Le CRF calcule la probabilité de la séquence de sortie Y étant donnée la séquence d'entrée X de la manière suivante : Dans notre cas, X correspond à la sortie de traduction automatique, et Y représente les étiquettes associées aux mots. Chaque élément y_i $(i = \overline{1..N})$ se voit attribuer une valeur dans l'ensemble binaire $Y^N = Good, Bad$. La probabilité de la séquence Y étant donné X est :

$$p_\theta(Y|X) = \frac{1}{Z_\theta(X)} exp\left\{\sum_{k=1}^{K} \theta_k F_k(X,Y)\right\} \tag{5.1}$$

où $F_k(X,Y) = \sum_{t=1}^{T} f_k(y_{t-1}, y_t, x_t)$; $\{f_k\}\,(k = \overline{1,K})$ est un ensemble de fonctions caractéristiques ; $\{\theta_k\}\,(k = \overline{1,K})$ sont les valeurs des paramètres associés ; et $Z_\theta(x)$ est une fonction de normalisation. Le lecteur trouvera plus de détails dans [Lavergne et al., 2010]. Les traits extraits du système de traduction sont relativement classiques et se rapprochent de ceux que l'on peut utiliser pour la parole, nous nous sommes par ailleurs appuyés sur les travaux de [Ueffing et al., 2003, Blatz et al., 2004, Bicici, 2013, Han et al., 2013] en ce

qui concerne les traits :

- Des caractéristiques sur le mot cible et son contexte d'alignement.
- Des caractéristiques du mot source et son contexte d'alignement.
- Probabilité *a posteriori* du mot cible et comportements de repli du modèle de langue [Mauclair, 2006].
- POS (Part-Of-Speech) : du mot cible et source.
- Des caractéristiques lexicales indiquant si le mot est un : mot d'arrêt, symbole de ponctuation, nom propre ou numérique.
- Des caractéristiques de la topologie du graphe.
- L'étiquette constitutive du mot et sa profondeur dans l'arbre issues de l'arbre constitutif produit par l'analyseur syntaxique Berkeley [Petrov and Klein, 2007].
- Le nombre de sens de chaque mot compte tenu de son POS (pour le mot cible).

Ces traits sont décrits exhaustivement dans [Luong et al., 2017]. Une fois extraits, ils servent à entraîner un modèle CRF qui étiquette les sorties de notre système. La figure 5.1 montre le résultat de cette annotation sur une phrase.

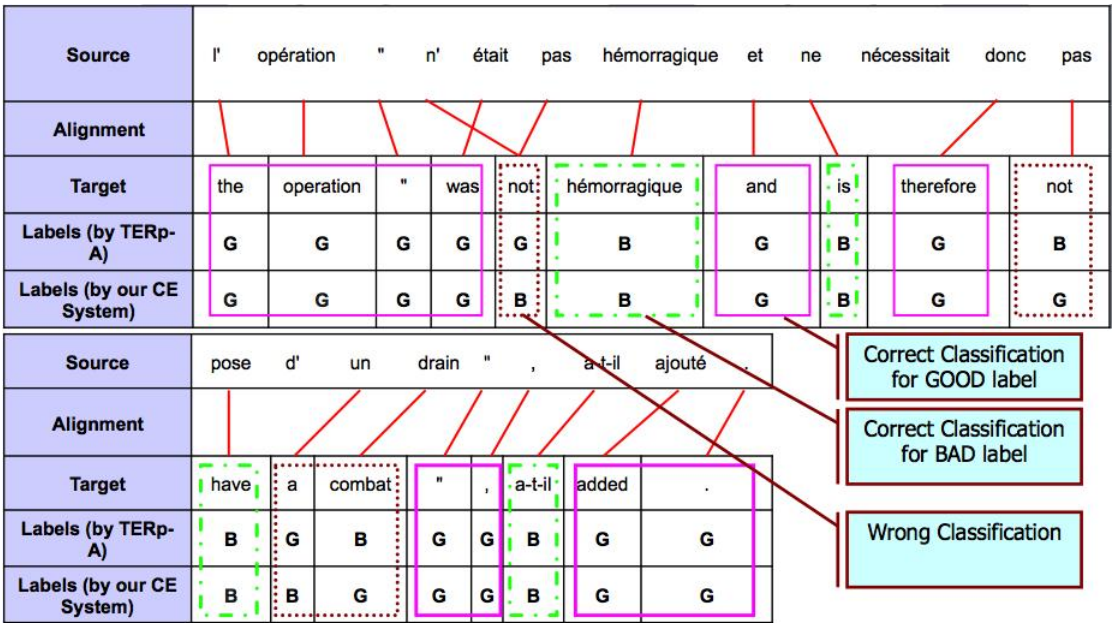


FIGURE 5.1 – Attribution d'étiquettes à partir de notre système de mesures de confiance. Figure extraite de [Luong et al., 2017].

5.1.2 Réévaluation des N-meilleures hypothèses

Les étiquettes étant associées au niveau du mot, nous devons les calculer au niveau de la phrase afin de les intégrer aux scores existants du décodeur *Moses* [Koehn et al., 2007b]. Six nouveaux scores sont proposés ; ils comprennent :

- Le rapport entre le nombre de mots corrects et le nombre total de mots → 1 score.
- Le rapport entre le nombre de noms/verbes corrects et le nombre total de noms/verbes → 2 scores.
- Le rapport entre le nombre de N-séquences correctes de mots consécutifs et le nombre total de séquences de mots consécutifs pour n=2, n=3 et n=4 → 3 scores.

Ces scores sont ensuite intégrés dans les N-meilleures hypothèses et permettent de réajuster leurs rangs. Les résultats des expériences avec réordonnancement sont présentés dans la section 5.1.3 avec les résultats des autres approches.

5.1.3 Réévaluation de graphe avec des mesures de confiance

Dans ce cas, les scores évalués sur les N-meilleures hypothèses vont être directement injectés au sein du graphe de décodage SG. Le décodeur génère N-meilleures hypothèses $T = \{T_1, T_2, ..., T_N\}$ lors de la première passe. En utilisant le système d'estimation de mesures de confiance, nous sommes en mesure d'attribuer au *j-ème* mot de l'hypothèse T_i, désigné par t_{ij}, une étiquette c_{ij} *"G"* (*Good* : aucune erreur de traduction), *"B"* (*Bad* : doit être édité)), ainsi que les probabilités de confiance $P_{ij}(G), P_{ij}(B)$ que nous simplifierons en $P(B)$ et $P(G)$, où $P(B) = 1 - P(G)$. La seconde passe est effectuée en considérant chaque mot t_{ij} ainsi que ses étiquettes et scores $c_{ij}, P(G), P(B)$. L'idée est que si t_{ij} est une bonne traduction (c_{ij} = "G" ou $P(G) \approx 1$), toutes les hypothèses $H_k \in SG$ qui la contiennent devraient être récompensées. Au contraire, celles contenant une mauvaise traduction devraient être pénalisées : nous proposons ainsi $reward(t_{ij})$ et $penalty(t_{ij})$ les scores de récompense ou de pénalité de t_{ij}. Le nouveau coût de transition de H_k après avoir été mis à jour est alors défini par :

$$transition'(H_k) = transition(H_k) + \begin{cases} reward(t_{ij}) & \text{if } t_{ij} = good \\ penalty(t_{ij}) & \text{if } otherwise \end{cases} \tag{5.2}$$

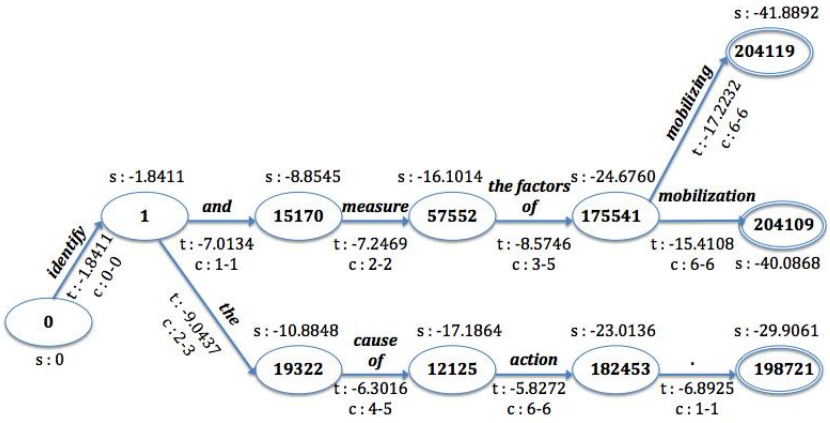


FIGURE 5.2 – Graphe avant ré-évaluation. Figure extraite de [Luong et al., 2017].

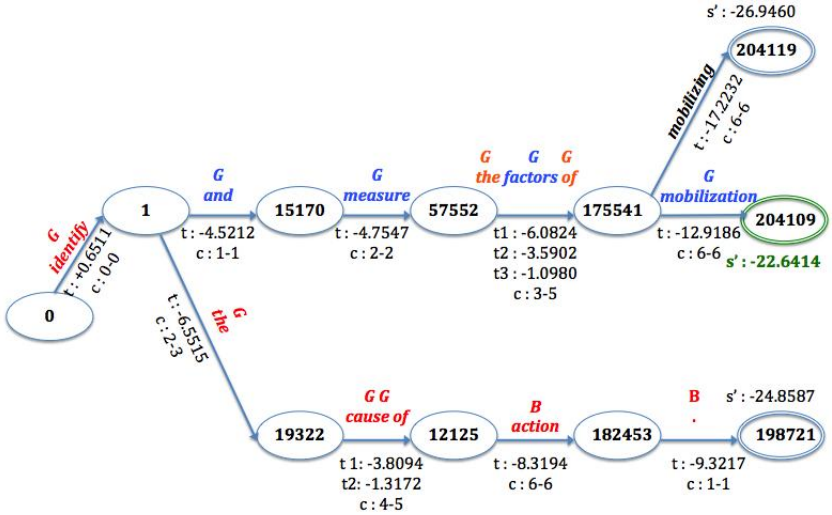


FIGURE 5.3 – Graphe avec ré-évaluation. La première passe est représentée en rouge, tandis que la seconde est en bleu. Le nouveau meilleur chemin est mis en évidence en vert. Figure extraite de [Luong et al., 2017].

La mise à jour se termine lorsque tous les mots de la N-meilleure liste ont été évalués. Nous recalculons alors le nouveau score des hypothèses complètes. La passe de re-décodage réorganise les hypothèses du graphe : plus elles contiennent de mots *"G"*, plus le coût sera réduit et, par conséquent, mieux elles seront classées.

Le détail de la ré-évaluation du graphe est présenté dans l'algorithme 1.

Algorithm 1 Algorithme de ré-évaluation du graphe

Input : $SG = \{H_k\}$, $T = \{T_1, T_2, ..., T_N\}$, $C = \{c_{ij}\}$
Output : $T^{'} = \{T_1^{'}, T_2^{'}, ..., T_N^{'}\}$

{**Step 1 : Update the Search Graph**}
$Processed \leftarrow \emptyset$
for T_i in T **do**
  for t_{ij} in T_i **do**
    $p_{ij} \leftarrow$ position of the source words aligned to t_{ij}
    if $(t_{ij}, p_{ij}) \in Processed$ **then**
      continue ; {ignore if t_{ij} appeared in the previous sentences}
    end if
    $Hypos \leftarrow \{H_k \in SG |\ out(H_k) \ni t_{ij}\}$
    if $(c_{ij} =$ "$Good$"$)$ **then**
      for H_k in $Hypos$ **do**
        $transition(H_k) \leftarrow transition(H_k) + reward(t_{ij})$ {reward hypothesis}
      end for
    else
      for H_k in $Hypos$ **do**
        $transition(H_k) \leftarrow transition(H_k) + penalty(t_{ij})$ {penalize hypothesis}
      end for
    end if
    $Processed \leftarrow Processed \cup \{(t_{ij}, p_{ij})\}$
  end for
end for
{**Step 2 : Trace back to re-compute the score for all complete hypotheses**}
for H_k in $Final$ (Set of complete hypotheses) **do**
  $score(H_k) \leftarrow 0$
  while $H_k \neq$ initial hypothesis **do**
    $score(H_k) \leftarrow score(H_k) + transition(H_k)$
    $H_k \leftarrow pre(H_k)$
  end while
end for
{**Step 3 : Select N cheapest hypotheses and output the new list** $T^{'}$}

Les figures 5.2 et 5.3 présentent les graphes avant et après décodage.

Expériences

Corpus d'apprentissage et de test

Nous avons traduit en anglais, avec un système de TA basé sur *Moses*, un ensemble de 10881 phrases françaises. Ensuite, des traducteurs humains ont été invités à corriger les sorties de TA, pour obtenir les post-éditions. Plus de

Systems	MERT			MIRA		
	BLEU	TER	TERp-A	BLEU	TER	TERp-A
BL	52.31	0.2905	0.3058	50.69	0.3087	0.3036
BL+OR	**58.10**	**0.2551**	**0.2544**	**55.41**	**0.2778**	**0.2682**
BL+WCE	52.77	0.2891	0.3025	51.01	0.3055	0.3012
WCE + 25%	53.45	0.2866	0.2903	51.33	0.3010	0.2987
WCE + 50%	55.77	0.2730	0.2745	53.63	0.2933	0.2903
WCE + 75%	56.40	0.2687	0.2669	54.35	0.2848	0.2822
Oracle BLEU score	**BLEU=60.48**					

TABLE 5.1 – Résultats avec ré-ordonnancement des hypothèses. WCE+25, 50 et 75 correspondent à un pourcentage de mesures oracles afin de voir le potentiel de la méthode en fonction de la qualité des mesures de confiance. Résultats extraits de [Luong et al., 2017].

détails sur ce corpus post-édité peuvent être trouvés dans [Potet et al., 2012]. L'ensemble de triplets (source, hypothèse, post-édition) a ensuite été divisé en un ensemble d'entraînement (10000 premiers triplets) et de test (881 triplets restants). Le classificateur WCE a été entraîné sur les hypothèses *1-best* de l'ensemble d'entraînement.

Résultats

La première approche consiste à réordonner les N-meilleures hypothèses en s'appuyant sur la méthode décrite dans la section 5.1.2 ; les résultats sont présentés dans la table 5.1.

Le ré-ordonnancement à l'aide des mesures de confiance montre un gain sur l'ensemble des mesures BLEU, TER et TERp. Par ailleurs, nous montrons le gain potentiel à l'aide de mesures oracles.

La seconde approche est basée sur la réévaluation de graphe telle que présentée dans la section 5.1.3. Les résultats de cette approche sont présentés dans la table 5.2

Ces résultats sont intéressants car ils montrent que réévaluer le graphe offre un potentiel plus grand de corrections. Dans notre cas, cette méthode s'apparente à une seconde passe telle qu'il en existe dans le domaine de la RAP.

Systems	Performance			Comparison to BL			p-value
	BLEU ↑	TER ↓	TERp-A ↓	B (%)	E (%)	W (%)	
BL	52.31	0.2905	0.3058	-	-	-	-
BL+WCE(1a)	**53.80**	**0.2876**	**0.2922**	28.72	57.43	13.85	0.00
BL+OR(1a)	**60.18**	**0.2298**	**0.2264**	62.52	24.36	13.12	-
Oracle BLEU = 66.48 (from SG)							

TABLE 5.2 – Résultats avec réévaluation du graphe : baseline (BL), BL+mesures de confiance (WCE), BL+oracle (OF). Résultats extraits de [Luong et al., 2017].

5.2 Désambiguïsation et traduction automatique

La thèse de Loïc Vial a principalement porté sur la désambiguïsation lexicale (DL), puis a abordé l'intégration de la DL au sein des systèmes de TA. Nous avons proposé des méthodes état-de-l'art, que je ne détaillerai pas dans ce manuscrit, pour répondre à une tâche de désambiguïsation lexicale et nous avons souhaité voir si cette information pouvait avoir un impact dans le cadre d'un système de traduction automatique. De mon point de vue, il s'agit d'une sorte de mesure de confiance directement extraite à partir du texte et de son contexte.

Nous avons ainsi proposé différentes méthodes pour injecter les informations de désambiguïsation lexicale au sein d'un système de TA. Nous avons également été précurseurs dans l'introduction de vecteurs contextualisés de type BERT [Devlin et al., 2019] au sein d'un système de TA [Vial et al., 2019d, Vial, 2020].

Plusieurs méthodes ont été proposées :

1. Ajouter l'information de désambiguïsation comme un trait supplémentaire. Cette information de désambiguïsation a été proposée dans les travaux de Loïc Vial [Vial et al., 2017b, Vial et al., 2019a]
2. Un apprentissage multi-tâche permettant d'apprendre à traduire et désambiguïser
3. Utiliser les vecteurs contextuels directement en entrée

Dans la première méthode, nous avons concaténé les informations "vecteur de sens" et "vecteur de mot" en un seul vecteur qui est pris en entrée du système de TA. Cette méthode est illustrée dans la figure 5.4.

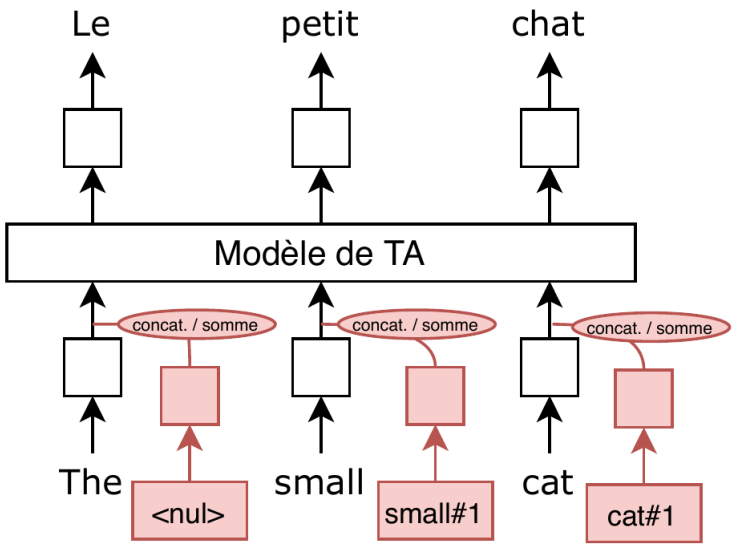


FIGURE 5.4 – Ajout des sens en tant qu'information discrète supplémentaire pour un système de TA neuronal. Figure extraite de [Vial, 2020].

La seconde méthode vise à guider l'apprentissage par le sens. Pour cela nous nous appuyons sur un apprentissage multi-tâche. Lorsque les tâches sont proches, il a été montré dans plusieurs travaux que cela pouvait être bénéfique pour la tâche principale. Cette méthode est illustrée dans la figure 5.5.

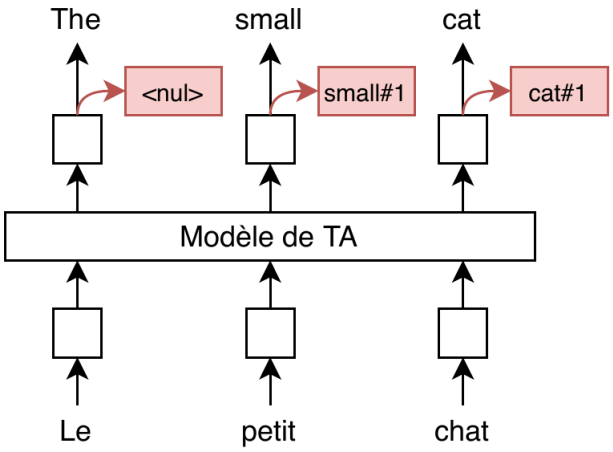


FIGURE 5.5 – Apprentissage guidé par le sens via un apprentissage multi-tâche. Figure extraite de [Vial, 2020].

La troisième méthode est semblable à la première, mais au lieu de détecter le sens séparément nous nous appuyons directement sur un modèle qui devrait pouvoir capturer cette information, à savoir BERT. Nous n'avons plus besoin

d'apprendre nos vecteurs de sens et nous utilisons les vecteurs contextuels comme entrée du système de TA. Cette méthode est illustrée dans la figure 5.6.

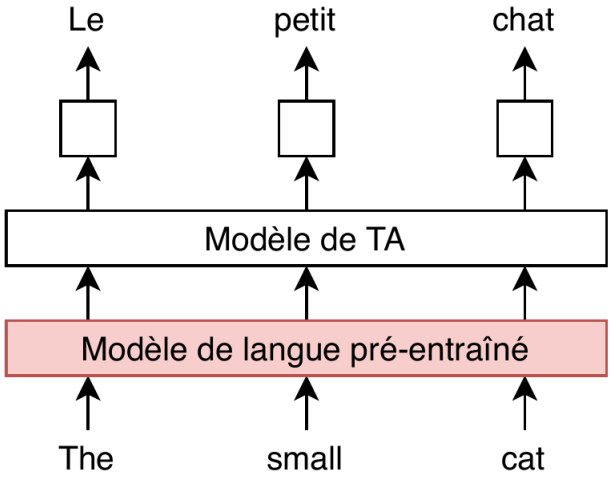


FIGURE 5.6 – Apprentissage exploitant directement les vecteurs contextuels de type BERT. Figure extraite de [Vial, 2020].

Résultats

Le système de TA implémenté s'inspire d'une architecture Transformer [Vaswani et al., 2017] avec encodeur et décodeur. Afin d'évaluer l'impact de nos méthodes en regard d'un système de référence connu, nous utilisons la tâche de traduction anglais→allemand issue de IWSLT 2014. Les données d'entraînement contiennent 171721 phrases parallèles anglais→allemand, correspondant à des transcriptions des conférences TED. Le corpus de test est constitué de 6750 phrases également issues des conférences TED. Au niveau des traductions, nous avons d'abord expérimenté la traduction anglais→allemand afin d'évaluer les méthodes 1 et 3 (en effet, elles consistent à ajouter des informations de sens sur la langue source, et nous n'avons qu'un système pour l'anglais). En ce qui concerne la méthode 2, où la langue cible est désambiguïsée, nous inversons les langues pour traduire de l'allemand vers anglais. L'annotation en sens est effectuée avec le système présenté dans [Vial et al., 2019a]. Les résultats de ces expériences sont rassemblés dans la table 5.3. Les méthodes d'intégration des sens en entrée améliorent dans tous les cas les performances. Cependant, l'approche multi-tâche dégrade les résultats : ce constat nous a semblé surprenant ; nous en avons déduit que les tâches de traduction et de désambiguïsation étaient trop disjointes pour que leur association s'avère pertinente. La méthode la plus performante se base sur BERT.

	Système	BLEU (%) ↑	TER (%) ↓	Meteor (%) ↑
anglais → allemand	⋆Bahdanau et al. (2015)	25,0	-	-
	⋆Gehring et al. (2017)	26,7	-	-
	⋆Vaswani et al. (2017)	**28,1**	-	-
	Elbayad et al. (2018)	27,2	-	-
	Mot → mot (référence)	26,8	53,43	47,10
	Mot + sens (somme) → mot	27,1	53,04	47,40
	Mot + sens (concat.) → mot	27,4	52,40	47,58
	BERT → mot	**29,5**	**49,64**	**49,04**
allemand → anglais	⋆Bahdanau et al. (2015)	29,9	-	-
	⋆Gehring et al. (2017)	32,3	-	-
	⋆Vaswani et al. (2017)	**34,4**	-	-
	Elbayad et al. (2018)	33,9	-	-
	Mot → mot (référence)	**30,3**	**47,25**	**34,27**
	Mot → mot + sens	30,1	47,58	34,18

TABLE 5.3 – Table extraite de la thèse de Loïc Vial : Résultats de nos méthodes d'intégration de la désambiguïsation Lexicale (DL) dans un système de TA sur la tâche de traduction allemand - anglais de IWSLT 2014. Les systèmes préfixés par une étoile sont des ré-implémentations issues de l'article [Elbayad et al., 2020a]. Les meilleurs résultats par tâche apparaissent en gras pour les systèmes de référence ainsi que pour nos propres systèmes. Les flèches (↑↓) indiquent le sens d'amélioration des scores. Les tirets (-) remplacent les scores non fournis par les auteurs. Table extraite de [Vial, 2020].

Enfin nous avons essayé de déterminer si l'approche BERT pouvait être améliorée en lui adjoignant une information experte liée à notre système de désambiguïsation. Pour cela nous avons concaténé les vecteurs contextuels avec nos vecteurs de sens. Les résultats attachés à cette approche sont présentés dans la table 5.4. Nous observons que, effectivement, BERT est perfectible et qu'une information experte liée au sens induit un gain de performance.

BERT	-	29,5	49,64	49,04
BERT	128	29,8	49,60	49,38
BERT	256	29,7	49,46	49,33
BERT	512	**29,9**	**49,40**	**49,49**
BERT	1024	29,7	49,44	49,17

TABLE 5.4 – Le système avec le tiret n'inclut aucun vecteur de sens. Ensuite, nous concaténons des vecteurs de sens de tailles différentes avec les vecteurs contextuels BERT. Les résultats sont des scores BLEU, TER et Meteor. Table extraite de [Vial, 2020].

5.3 Conclusion

Dans ce chapitre nous avons abordé les mesures de confiance spécifiques à la traduction automatique. La problématique est un peu similaire à celle du domaine de la parole, mais en s'abstrayant du signal. Elle est par ailleurs, de mon point de vue, complexifiée par la multiplicité des modèles qui sont utilisés dans les systèmes probabilistes de type *phrase-based* et par la difficulté intrinsèquement liée tant à la traduction automatique qu'à son évaluation : dans le cadre de la transcription de parole, l'évaluation est relativement aisée et formelle ; en ce qui concerne la traduction, le BLEU est une métrique qui comporte certains biais. Cependant, nous avons établi que des mesures de confiance étaient aptes à orienter un second décodage en forçant le système à emprunter des chemins alternatifs dans le cadre d'un système probabiliste. Nous avons montré dans la thèse de Loïc Vial [Vial, 2020], qu'un système de TA neuronal peut également tirer parti d'informations externes : l'ajout d'informations sur le sens des mots est capable d'influer sur le système. Ce type d'approche est maintenant généralisé avec les représentations vectorielles contextuelles de type BERT [Devlin et al., 2019, Le et al., 2020] qui permettent d'encoder, parmi d'autres informations, des notions de sens. Il s'avère par ailleurs, comme nous l'avons expérimenté, que l'on peut parfaire les vecteurs issus de BERT en les concaténant avec des informations expertes. Le chapitre suivant abordera la continuité des travaux de Ngoc Quang Luong, à savoir la traduction automatique de la parole et la jointure des mesures de confiance issues des deux modalités.

Chapitre 6

Mesures de confiance et traduction automatique de la parole

Encadrements : Ngoc Tien Le, Ngoc Quang Luong

Sélection de publications relatives à ce chapitre :
N-T. Le, **B. Lecouteux**, L. Besacier. Automatic quality estimation for speech translation using joint ASR and MT features. Machine Translation 2018.

NT. Le, **B. Lecouteux**, L. Besacier. Disentangling ASR and MT Errors in Speech Translation. MT Summit 2017.

NT. Le, C. Servan, **B. Lecouteux**, L. Besacier. Better Evaluation of ASR in Speech Translation Context Using Word Embeddings. InterSpeech 2016.

NT. Le, **B. Lecouteux**, L. Besacier. Joint ASR and MT Features for Quality Estimation in Spoken Language Translation. IWSLT 2016.

C. Servan, NT. Le, NQ. Luong, **B. Lecouteux**, L. Besacier. An Open Source Toolkit for Word-level Confidence Estimation in Machine Translation. IWSLT 2015.

Suite à la thèse de Ngoc Quang Luong portant sur les mesures de confiance pour la traduction automatique, nous avons étendu ses travaux en vue d'améliorer les systèmes de traduction automatique de la parole. Les systèmes évoqués sont de type "cascadé" où nous souhaitions réduire au maximum la propagation d'erreurs du système de reconnaissance automatique de la parole au système de traduction automatique. Un premier obstacle a posé question : l'évaluation des erreurs issues du SRAP. En effet, dans ce type d'application le taux d'erreurs mots n'est pas la mesure la plus pertinente, car il ne considère pas du tout les aspects sémantiques qui sont très importants pour un système de traduction automatique. Une autre difficulté qu'il fallait surmonter était la faible disponibilité de corpus de traduction automatique de la parole qui permette d'évaluer les mesures de confiances aux différents niveaux de la chaîne de traduction de la parole. La thèse [Le, 2018] de Ngoc Tien Le a porté sur les mesures de confiance pour la traduction automatique de la parole. Ses travaux [Servan et al., 2015, Le et al., 2016a, Le et al., 2016b, Le et al., 2017, Le et al., 2018] se sont ainsi attachés aux différentes manières de combiner ou joindre des mesures de confiance issues du SRAP et du système de TA.

6.1 Évaluation des sorties SRAP pour la TA

Dans le domaine de la traduction de la parole, la métrique du taux d'erreurs mots pour évaluer l'impact du module de RAP sur l'ensemble du pipeline "parole→traduction" est souvent remise en question. Cet aspect a soulevé plusieurs travaux où les auteurs ont tenté de proposer une meilleure évaluation dans ce scénario. [Dixon et al., 2011] a démontré que des taux de WER sous-optimaux pouvaient donner des scores BLEU comparables tout en accélérant les vitesses de décodage. D'autres travaux comme ceux de [Bechet et al., 2015] ont analysé les segments erronés issus du SRAP ayant un impact élevé sur les performances et ont démontré que la suppression de ces segments avant la traduction améliorait le résultat final.

Nos travaux se sont appuyés sur ceux de [Vilar et al., 2006] qui a observé que de nombreuses erreurs de substitution sont liées à de légers changements morphologiques (pluriel/singulier par exemple), ce qui limite l'impact lors de la traduction. Ainsi, la métrique WER est sous-optimale pour évaluer le module de reconnaissance automatique de la parole quand on doit la traduire.

Nous avons donc proposé une extension du WER pour évaluer les erreurs de substitution en fonction de leur contexte. La métrique idéale doit moins pénaliser les changements morphologiques qui ont un impact limité et considérer

positivement une sémantique conservée.

Nous avons créé un corpus spécifique français→anglais afin d'évaluer notre méthode et nous avons évalué la corrélation de notre nouvelle métrique avec les performances du système de traduction.

Concernant la traduction automatique, [Gupta et al., 2015] a proposé une métrique qui représente à la fois la référence et sa traduction à l'aide d'un réseau de neurones de type (*Long Short Term Memory LSTM*) qui prédit le score de similarité. [Vela and Tan, 2015] a utilisé des représentations vectorielles de documents pour prédire leur adéquation avec la traduction automatique. Ces travaux reposent tous deux sur l'apprentissage de la métrique elle-même, ce qui limite leur portabilité pour l'évaluation dans d'autres conditions.

Notre proposition est relativement simple : effectuer les comparaisons source/cible à l'aide de plongements de mots qui seront plus sensibles aux aspects sémantiques et moins aux aspects morphologiques. Dans le cas du WER, la mesure utilisée est une distance de Levenshtein où la substitution est pénalisée de manière fixe comme le montre la figure 6.1.

		un	nord	westphalie	un	d'	engagement	parmi	de	nation	souveraine
		0	1	2	3	4	5	6	7	8	9
un		1	1	2	3	4	5	6	7	8	9
ordre		2	2	2	3	4	5	6	7	8	9
westphalien		3	3	3	3	3	4	5	6	7	8
d'		4	4	4	4	4	4	5	6	7	8
engagements		5	5	5	5	5	5	4	5	6	7
parmi		6	6	6	6	6	6	5	5	6	7
des		7	7	7	7	7	7	6	6	6	7
nations		8	8	8	8	8	8	7	7	7	7
souveraines											
Alignment:		A	I	S	S	A	S	A	S	S	S
Cost:		0	1	1	1	0	1	0	1	1	1

TABLE 6.1 – Exemple de distance de Levenshtein. Table extraite de [Le et al., 2016b].

Nous avons proposé deux méthodes :

1. Remplacer le score de substitution par une distance cosine entre le mot de la référence et celui de l'hypothèse (WER-E) comme présenté dans la table 6.2 : cela nous donne un nouveau score.
2. Utiliser la première méthode et recalculer l'alignement de Levenshtein qui a pu être impacté (WER-S), ce qui permet de trouver un meilleur chemin comme montré dans la table 6.3.

	un	nord	westphalie	un	d'	engagement	parmi	de	nation	souveraine
un	0	1	2	3	4	5	6	7	8	9
ordre	1	1.01	2.07	2.93	4.15	4.89	6.07	7.03	8.05	9.01
westphalien	2	1.79	1.73	2.83	3.93	5.38	5.80	6.90	7.75	8.85
d'	3	3.05	2.97	2.21	2.83	3.83	4.83	5.83	6.83	7.83
engagements	4	3.94	4.02	4.15	3.41	3.30	5.01	5.91	6.92	7.81
parmi	5	4.77	4.80	5.13	5.15	4.61	3.30	4.30	5.30	6.30
des	6	6.04	5.85	5.80	5.61	6.24	4.30	3.64	5.49	6.12
nations	7	6.87	6.83	6.77	6.85	6.55	5.30	5.26	4.42	6.43
souveraines	8	7.92	7.71	7.99	7.71	7.82	6.30	6.15	6.10	**4.85**
Alignment:		A	I	S	S	A	S	A	S	S	S
Cost:		0	1	1.07	0.75	0	0.47	0	0.35	0.78	0.43

TABLE 6.2 – Exemple de distance de Levenshtein en modifiant le score de substitution par une distance *cosine*. Table extraite de [Le et al., 2016b].

	un	nord	westphalie	un	d'	engagement	parmi	de	nation	souveraine
un	0	1	2	3	4	5	6	7	8	9
ordre	1	1.01	2.01	2.93	3.93	4.89	5.89	6.89	7.89	8.89
westphalien	2	1.79	1.74	2.74	3.74	4.74	5.74	6.72	7.61	8.61
d'	3	2.79	2.74	2.21	2.74	3.74	4.74	5.74	6.74	7.74
engagements	4	3.79	3.74	3.21	3.42	3.21	4.21	5.21	6.21	7.21
parmi	5	4.77	4.65	4.21	4.21	4.21	3.21	4.21	5.21	6.21
des	6	5.77	5.65	5.21	4.68	5.21	4.21	3.55	4.55	5.55
nations	7	6.77	6.57	6.21	5.68	5.63	5.21	4.55	4.34	5.34
souveraines	8	7.77	7.57	7.21	6.68	6.63	6.21	5.55	5.34	4.76
Alignment:		A	S	S	I	A	S	A	S	S	S
Cost:		0	1.01	0.73	1	0	0.47	0	0.35	0.78	0.43

TABLE 6.3 – Exemple de distance de Levenshtein en modifiant le score de substitution par une distance *cosine* et en recalculant les chemins. Table extraite de [Le et al., 2016b].

Ces méthodes vont être utilisées pour optimiser l'entraînement d'un système de traduction automatique de la parole. La section suivante présente les résultats obtenus grâce à cette nouvelle métrique.

Résultats

Nous avons regardé si les métriques WER-E et WER-S sont mieux corrélées avec les performances de la traduction de la parole et nous avons évalué un système de TA optimisé avec ces métriques. Les résultats sont présentés dans

la table 6.4. Nous constatons une meilleure corrélation des métriques proposées (WER-E et WER-S) et une amélioration des performances de TA, par rapport à la métrique classique du WER. Nous observons également que toutes les métriques WER sont mieux corrélées avec METEOR (lui-même connu pour être mieux corrélé avec les jugements humains), tandis qu'elles sont moins corrélées avec BLEU.

Tasks	*metrics*	Pearson Correlation		
		WER	WER-E	WER-S
dev	*TER*	0.732	0.767	**0.773**
	BLEU	-0.677	-0.708	**-0.710**
	METEOR	-0.753	**-0.799**	-0.797
tst	*TER*	**0.457**	**0.457**	0.441
	BLEU	-0.624	**-0.661**	-0.606
	METEOR	-0.672	**-0.692**	-0.678

Tasks	*metrics*	ASR optimized with WER	ASR optimized with WER-E
dev	*TER*	55.64	**55.52**
	BLEU	30.81	**30.84**
	METEOR	**34.02**	34.00
test	*TER*	58.71	**58.56**
	BLEU	34.27	**34.38**
	METEOR	**34.27**	34.26

TABLE 6.4 – Tests de corrélation des différentes mesures et résultats du système de TA après optimisation avec les différentes mesures. Table extraite de [Le et al., 2016b].

Les résultats de traduction sont quant à eux peu convaincants : nous observons de faibles gains en terme de TER et BLEU et aucune pour METEOR. Notre explication réside dans le manque de paramètres permettant d'ajuster le SRAP. De plus, les métriques d'évaluation de la traduction sont imparfaites. Cependant, des expériences "Oracles" qui ne sont pas détaillées ici montrent que nos métriques ont malgré tout réussi à trouver les meilleures hypothèses à traduire en comparaison d'un WER classique sur les 1000-meilleures phrases.

6.2 Création d'un corpus spécifique

Afin de créer un corpus (français parlé)→(anglais écrit), nous sommes partis d'un travail effectué par Laurent Besacier [Potet et al., 2010, Potet et al., 2012] : un système de traduction a été utilisé pour obtenir les hypothèses de traduction de 10881 phrases issues de plusieurs corpus de nouvelles provenant des campagnes d'évaluation WMT (Workshop on Machine Translation) (de 2006 à 2010). Les post-éditions ont été obtenues auprès de traducteurs non professionnels en utilisant une plateforme de *crowdsourcing*. Un sous-ensemble (311 phrases) de ces post-éditions a été évalué par un traducteur professionnel : 87,1% de l'ensemble des post-éditions ont été jugées comme améliorant l'hypothèse du système de traduction automatique.

L'étiquetage des mots pour les estimations de confiance (*Good* ou *Bad*) a été effectué à à l'aide de l'outil TERp [Snover et al., 2008]. Chaque mot

ou phrase de l'hypothèse est aligné sur un mot ou une phrase de la post-édition avec différentes étiquettes d'édition : "I" (insertions), "S" (substitutions), "T" (correspondances de racines), "Y" (correspondances de synonymes) et "P" (substitutions au niveau du segment). L'absence de symbole indique une correspondance exacte et sera remplacée par "E" par la suite. Nous ne considérons pas les mots marqués d'un "D" (suppressions) puisqu'ils n'apparaissent que dans la référence. Cependant, par la suite, afin d'entraîner un classificateur binaire (*good*/*bad*) nous re-catégorisons l'ensemble de 6 étiquettes : (E, T et Y) appartiennent à la catégorie *good* (G), tandis que (S, P et I) appartiennent à la catégorie *bad* (B).

Afin d'avoir les correspondances orales de notre corpus source (Français), nous avons réalisé des enregistrements avec des volontaires : les ensembles *dev* et *tst* du corpus ont été enregistrés par des locuteurs natifs français. Chaque phrase a été prononcée par 3 locuteurs, ce qui donne respectivement 2643 enregistrements pour le *dev* et 4050 pour le *tst*. Au final, pour chaque énoncé, un quintuplet contenant : la sortie du SRAP (f_{hyp}), la transcription verbatim (f_{ref}), la sortie de la traduction du texte source vers l'anglais ($e_{hyp_{mt}}$), la sortie de la traduction de la parole ($e_{hyp_{slt}}$) et la post-édition de la traduction (e_{ref}), a été mis à disposition. Ce corpus est disponible sur le dépôt `https://github.com/besacier/WCE-SLT-LIG/`. La table 6.5 détaille ce corpus. La longueur totale des corpus de parole *dev* et *tst* obtenus est de 16h52.

Corpus	**#phrases**	**#enreg.**	**#locuteurs**	**Durée**
dev	881	2643	15 (9 femmes + 6 hommes)	5h51
tst	1350	4050	27 (11 femmes + 16 hommes)	11h01

TABLE 6.5 – Détails sur le corpus créé pour la traduction automatique de la parole [Le et al., 2018] .

Pour obtenir les transcriptions de la parole (f_{hyp}), nous avons construit un système de reconnaissance français basé sur *Kaldi* [Povey et al., 2011b]. Les modèles acoustiques ont été entraînés en utilisant les corpus ESTER, REPERE, ETAPE et BREF120. La transcription est ensuite réalisée en deux passes avec un système hybride HMM-DNN qui est détaillé dans [Le et al., 2018].

Nous proposons deux transcriptions : la première utilise un petit modèle de langage permettant un système RAP rapide (environ 2xRT), tandis que dans la seconde un grand modèle de langage est utilisé (environ 10xRT) lors d'une troisième passe.

Les résultats en termes de WER sont donnés dans la table 6.6. Les taux

d'erreurs peuvent sembler élevés pour de la parole lue. Ceci s'explique par le fait qu'il y a beaucoup d'entités nommées inconnues (et parfois difficiles à prononcer).

TABLE 6.6 – Performances de nos systèmes de reconnaissance de la parole sur le *dev* et le *tst* pour les deux configurations

Système	***dev***	***tst***
petit modèle de langue (SRAP1)	21.86%	17.37%
grand modèle de langue (SRAP2)	16.90%	12.50%

Ce corpus aura demandé une grande quantité de travail et sera par la suite utilisé durant toute la thèse de Ngoc Tien Le. La section suivante aborde les travaux qu'il a réalisés à l'aide de ce corpus.

6.3 Traduction de la parole et mesures de confiance

La thèse de Ngoc Tien Le a porté sur l'amélioration des systèmes de traduction automatique de la parole à l'aide de mesures de confiance jointes. La base de ses travaux s'appuient à la fois sur la thèse de Ngoc Quang Luong, les travaux de Laurent Besacier [Besacier et al., 2014] et le travail que nous avions réalisé autour des mesures de confiance pour la reconnaissance automatique de la parole. Pour la traduction automatique de la parole, les éléments suivants nous ont été nécessaires :

- Un corpus oral traduit : celui décrit dans la section précédente.
- Un système évaluant des mesures de confiance pour la reconnaissance automatique de la parole : ce système s'appuie sur celui présenté dans les sections 3.1 et 4.1.1. Nous avons cependant adapté ces travaux en utilisant un classifieur basé sur des CRF qui se sont avérés plus performants.
- Un système évaluant des mesures de confiance pour la traduction automatique : nous nous sommes appuyés sur les travaux de Ngoq Quang Luong présentés dans la section 5.1.1.
- Un système RAP : il a été construit de manière assez classique et est présenté dans la section précédente.
- Un système de TA : nous avons choisi de partir sur le couple français→anglais pour lequel nous possédions les différentes modalités (source, enregistrements de la source, cibles).

À partir de ces composants les mesures de confiance pour la traduction de la parole se définissent ainsi.

Étant donné un signal x_f dans la langue source, la traduction de la parole (*Speech Language Translation SLT*) consiste à trouver la séquence de la langue cible $\hat{e} = (e_1, e_2, ..., e_N)$ de sorte que

$$\hat{e} = \underset{e}{\operatorname{argmax}}\{p(e|x_f, f)\} \tag{6.1}$$

où $f = (f_1, f_2, ..., f_M)$ est la transcription de x_f.

Si nous estimons la confiance au niveau mot (WCE), cette information peut se représenter comme une séquence $q = (q_1, q_2, ..., q_N)$ avec $q_i \in \{good, bad\}$

L'évaluation des mesures de confiance pour la TA de la parole peut se définir alors comme suit :

$$\hat{e} = \underset{e}{\operatorname{argmax}} \sum_q p(e, q|x_f, f) \tag{6.2}$$

$$\hat{e} = \underset{e}{\operatorname{argmax}} \sum_q p(q|x_f, f, e) \times p(e|x_f, f) \tag{6.3}$$

$$\hat{e} \approx \underset{e}{\operatorname{argmax}}\{\max_q\{p(q|x_f, f, e) \times p(e|x_f, f)\}\} \tag{6.4}$$

Les probabilités SLT $p(e|x_f, f)$ et WCE $p(q|x_f, f, e)$ contribuent conjointement à trouver la meilleure traduction $\hat{e}$.

Le composant d'estimation de la confiance doit alors résoudre l'équation suivante :

$$\hat{q} = \underset{q}{\operatorname{argmax}}\{p_{SLT}(q|x_f, f, e)\} \tag{6.5}$$

où $q = (q_1, q_2, ..., q_N)$ est la séquence d'étiquettes de confiance pour la langue cible générée avec des CRF [Lafferty et al., 2001]. Pour cet apprentissage nous avons besoin de données d'apprentissage pour lesquelles un quadruplet (x_f, f, e, q) est disponible. Nous utiliserons pour cela le corpus décrit précédemment.

Comme les triplets (x_f, f, q) et (f, e, q) qui correspondent respectivement au :

- Discours transcrit automatiquement avec des références manuelles et des étiquettes issues du taux d'erreurs mots.
- Texte traduit automatiquement avec des post-éditions manuelles et des

étiquettes de confiance issues de TERp [Snover et al., 2008].

sont plus faciles à obtenir, nous proposons une autre approche pour calculer les mesures de confiance. Cette approche est décrite dans l'équation suivante :

$$\hat{q} = \underset{q}{\text{argmax}}\{p_{SRAP}(q|x_f, f)^{\alpha} \times p_{MT}(q|e, f)^{1-\alpha}\} \tag{6.6}$$

où α est un facteur d'échelle pour WCE_{SRAP} (qualité de la transcription) par rapport à WCE_{MT} (qualité de la traduction). $p_{SRAP}(q|x_f, f)$ correspond à l'estimation de la confiance dans la langue cible à partir de caractéristiques calculées sur la langue source (SRAP). Les deux approches : combinaison 6.6 et jointe 6.5 sont évaluées.

Résultats

La table 6.7 présente les résultats de performance de nos mesures de confiances en utilisant séparément les caractéristiques SRAP ou TA. Nous évaluons ainsi les performances des systèmes suivants :

- (WCE pour SRAP / SRAP) utilise les caractéristiques SRAP avec un classificateur CRF.
- (WCE pour SLT / TA) utilise uniquement les caractéristiques TA avec un classificateur CRF.
- (WCE pour SLT / SRAP) utilise uniquement les caractéristiques SRAP avec un classificateur CRF : nous prédisons la confiance de la sortie de TA en utilisant uniquement les caractéristiques de confiance SRAP. Les informations sur l'alignement des mots entre f_{hyp} et e_{hyp} sont utilisées pour projeter les scores WCE provenant du SRAP, vers la sortie SLT.

La table 6.8 présente les résultats de WCE utilisant les caractéristiques MT et SRAP. Nous évaluons les approches (*combinaison* et *jointe*) :

- La première combine les résultats de deux classificateurs SRAP et MT suivant l'équation 6.6.
- Le seconde entraîne un seul système de WCE suivant l'équation 6.5.

Il appert des résultats que les mesures de confiance jointes sont capables de dépasser les mesures de confiance combinées. Nous avons ensuite étudié la contribution de chaque caractéristique (SRAP ou TA) en appliquant une sélection de caractéristiques pour le classifieur joint. Nous choisissons l'algorithme SBS (*Sequential Backward Selection*) [Whitney, 1971], qui est un algorithme descendant partant d'un ensemble de caractéristiques noté Y_k (qui désigne

système :	**WCE pour SRAP**	**WCE pour TA**	**WCE pour SLT**
Paramètres :	SRAP $p(q\|x_f, f)$	TA $p(q\|f, e)$	SRAP $p_{SRAP}(q\|x_f, f)$ projetée sur e
F-mesure SRAP1	68.71%	64.69%	53.85%
F-mesure SRAP2	59.83%	64.48%	48.67%

TABLE 6.7 – Performance des mesures de confiance en utilisant différents paramètres sur le *tst*, le classifieur CRF est celui présenté dans notre article. SRAP1 et SRAP2 correspondent aux deux systèmes RAP que nous avons utilisés avec respectivement un petit modèle de langue et un grand modèle de langue [Servan et al., 2015].

tâche :	**WCE for SLT**	**WCE for SLT**
Approche :	combinaison SRAP+TA $p_{SRAP}(q\|x_f, f)^{\alpha} \times p_{MT}(q\|e, f)^{1-\alpha}$	Jointe SRAP/TA $p(q\|x_f, f, e)$
F-mesure SRAP1	58.07%	64.90%
F-mesure SRAP2	53.66%	64.17%

TABLE 6.8 – Performance des approches par combinaison ou jointe sur le *tst* et le *dev* ; nous observons que les mesures de confiance sont meilleures sur le système 1 : celui-ci étant plus mauvais, les mots erronés sont plus facilement détectés.

l'ensemble de toutes les caractéristiques) et qui élimine à chaque itération la caractéristique (x) maximisant la F-mesure moyenne, $MF(Y_k - x)$. L'algorithme 2 présente cette approche.

Les résultats de sélection disponibles en détail dans [Le et al., 2018] montrent que les caractéristiques issues de la TA sont les plus influentes, tandis que celles issues du SRAP peuvent apporter des informations complémentaires.

6.4 Conclusion

Dans ce chapitre nous avons présenté les travaux réalisés dans le cadre de la thèse de Ngoc Tien Le portant sur les mesures de confiance pour la traduction automatique de la parole. Nous avons ainsi proposé plusieurs contributions : la création d'un corpus dédié à la tâche, des outils permettant de générer des mesures de confiance pour les systèmes de RAP et de TA et des approches

Algorithm 2 Y_k est l'ensemble des paramètres et x le paramètre enlevé à chaque itération.

while size of $Y_k > 0$ **do**
  $maxval = 0$
  for $x \in Y_k$ **do**
    if $maxval < MF(Y_k - x)$ **then**
      $maxval \leftarrow MF(Y_k - x)$
      $worstfeat \leftarrow x$
    end if
  end for
  remove $worstfeat$ from Y_k
end while

combinées ou jointes pour générer des mesures de confiance. il en ressort que les mesures jointes surpassent toutes les autres approches. Ce résultat est extrêmement intéressant car dans le cadre de systèmes cascadés, il offre l'opportunité d'avoir un contrôle sur la propagation d'erreurs. C'est une approche que nous souhaiterons développer dans le cadre du projet PROPICTO présenté dans le chapitre 10. Le chapitre suivant présente les différentes campagnes d'évaluation et les principales réalisations industrielles auxquelles nous avons contribué.

Chapitre 7

Campagnes d'évaluation et réalisations industrielles

Encadrements : Solène Evain, Ngoc Tien Le, Ngoc Quang Luong, Loïc Vial, Lucia Ormaechea

Sélection de publications relatives à ce chapitre :

L. Ormachea Grijalba, P. Bouillon, J. Gerlach, **B. Lecouteux**, D. Schwab, H. Spechbach. Reconnaissance vocale du discours spontané pour le domaine médical. TLH 2021.

S. Evain, **B. Lecouteux**, F. Portet, I. Estève, M. Fabre. Towards Automatic Captioning of University Lectures for French Students who are Deaf. ACM SIGACCESS 2020.

M. Elbayad, H. Nguyen, F. Bougares, N. Tomashenko, A. Caubrière, **B. Lecouteux**, Y. Estève, L. Besacier. ON-TRAC Consortium for End-to-End and Simultaneous Speech Translation Challenge Tasks at IWSLT 2020. IWSLT 2020.

L. Vial, **B. Lecouteux**, D. Schwab, H. Le, L. Besacier. The LIG system for the English-Czech Text Translation Task of IWSLT 2019. IWSLT 2019.

NQ. Luong, L. Besacier, **B. Lecouteux**. LIG System for Word Level QE task at WMT14. WMT 2014

N.Q. Luong, **B. Lecouteux**, L. Besacier. LIG System for WMT13 QE Task : Investigating the Usefulness of Features in Word Confidence Estimation for MT. WMT 2013

L. Besacier, **B. Lecouteux**, M. Azouzi, N.Q. Luong. The LIG English to French Machine Translation System for IWSLT 2012

B. Lecouteux, L. Besacier, H. Blanchon : LIG English-French spoken language translation system for IWSLT 2011. IWSLT 2011

M. Potet, R. Rubino, **B. Lecouteux**, S. Huet, H. Blanchon, L. Besacier, F. Lefevre « The LIG/A Machine Translation System for WMT 2011

B. Lecouteux, L. Besacier, H. Blanchon : LIG English-French spoken language translation system for IWSLT 2011. IWSLT 2011

M. Potet, R. Rubino, **B. Lecouteux**, S. Huet, H. Blanchon, L. Besacier, F. Lefevre « The LIG/A Machine Translation System for WMT 2011

Depuis mon arrivée au sein de l'équipe GETALP, j'ai régulièrement été impliqué dans des campagnes d'évaluation [Potet et al., 2011, Lecouteux et al., 2011a, Besacier et al., 2012, Luong et al., 2013b, Luong et al., 2014b, Vial et al., 2019d, Elbayad et al., 2020b, Le et al., 2021] et des projets industriels. Ce chapitre aborde succinctement les campagnes et projets auxquels j'ai participé. Ils sont également récapitulés dans la figure 7.1.

7.1 Campagnes d'évaluation

Les campagnes d'évaluation sont importantes pour la visibilité de l'équipe et également pour la valorisation des travaux de nos doctorants. Elles permettent également de nous confronter aux équipes internationales, qu'elles viennent du secteur privé ou public. Nous ne participons pas systématiquement à toutes les campagnes d'évaluation : nous répondons aux sollicitations lorsqu'elles sont dans le focus de nos recherches. Nos participations se sont parfois faites avec peu de ressources et de temps, mais elles nous permettent à chaque fois de construire et proposer des systèmes à l'état-de-l'art. Nous présentons ici celles auxquelles nous avons participé. Chaque campagne est décrite très succinctement, tout

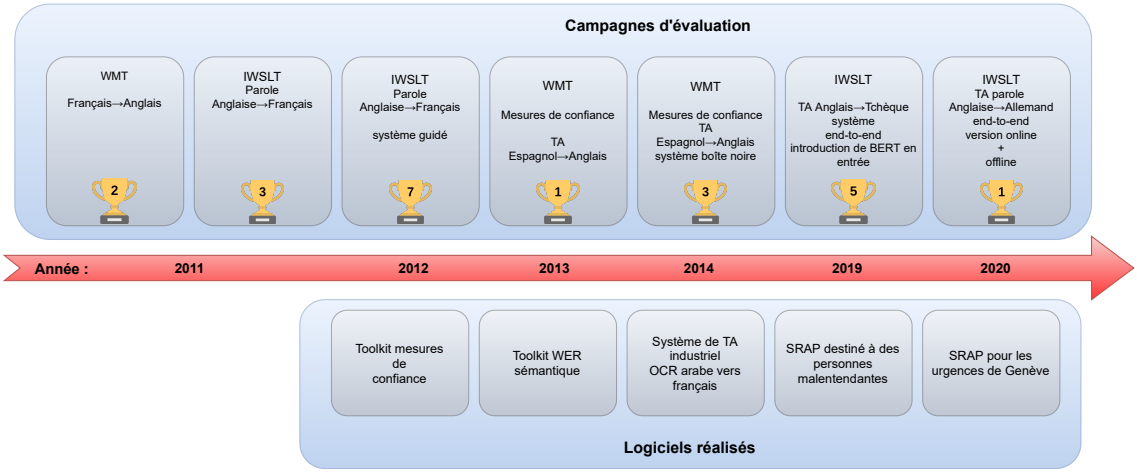


FIGURE 7.1 – Récapitulatif des différentes campagnes d'évaluation et logiciels auxquels nous avons contribué.

comme les résultats (nous ne détaillerons pas les autres participants) :

WMT 2011

Dès mon arrivée dans l'équipe GETALP, j'ai participé à ma première campagne d'évaluation liée à la traduction automatique de la parole en collaboration avec le laboratoire où je venais d'effectuer ma thèse (Laboratoire Informatique d'Avignon LIA). Nous avons participé à la tâche de traduction Français→Anglais. Chaque laboratoire a développé indépendamment son système et nous avons réalisé des fusions au niveau des graphes de recherche. Nous avons été classés seconds sur 18 participants. Les résultats sont présentés dans la table 7.1.

		LIG	LIA	LIG CNC	LIA CNC	LIG+LIA
case-insensitive BLEU	*test10*	27.6	29.3	28.1	29.4	29.7
	test11	28.5	29.4	28.5	29.3	29.9
case-sensitive BLEU	*test10*	26.1	28.4	27.0	28.4	28.7
	test11	26.9	28.4	27.5	28.4	28.8

TABLE 7.1 – Résultats sur la tâche français→anglais à la campagne d'évaluation WMT 2011 [Potet et al., 2011].

IWSLT 2011

Nous avons participé à la tâche (parole anglaise)→français où différentes sorties de SRAP étaient fournies. Nous avons proposé plusieurs approches pour tirer profit des multiples sorties du système de RAP. Les résultats expérimentaux montrent que la combinaison d'hypothèses de traduction obtenues à partir de plusieurs *1-best* du SRAP est la meilleure stratégie. Les résultats sont présentés dans la table 7.2. Nous avons terminé troisièmes sur cinq participants.

System.	bleu(p+c)	bleu(x)
LIG_P (Tst2011) source+target comb. (4201R)	0,2485	0,2598
LIG_C1 (Tst2011) source comb. (4201)	0,2453	0,2561
LIG_PostEval (Tst2011) Target comb (4201)	**0.2489**	**0.2599**

TABLE 7.2 – Résultats sur la tâche de traduction de la parole anglais-français à la campagne d'évaluation IWSLT 2011 [Lecouteux et al., 2011a].

IWSLT 2012

Lors de cette campagne, nous avons proposé une adaptation préliminaire du concept de mes précédents travaux sur le décodage guidé adapté à la traduction automatique. Cette méthode permet une combinaison efficace de systèmes de TA, en ré-évaluant le modèle log-linéaire au niveau des *N-best* en fonction de systèmes auxiliaires : la technique de base consiste essentiellement à guider la recherche en utilisant les sorties d'un ou plusieurs systèmes précédents. Les résultats montrent que cette approche permet une amélioration importante du score BLEU en utilisant *Google translate* pour guider notre système (un tel système a été présenté comme contrastif puisqu'il utilise un système de traduction automatique en ligne). Nous avons également essayé d'utiliser une mesure de confidence comme caractéristique log-linéaire supplémentaire, mais nous n'avons pu obtenir aucune amélioration avec cette technique. Cependant, le système primaire proposé était bien meilleur (+ 3 points de BLEU sur tst2010) par rapport à notre système présenté à IWSLT 2011. Les résultats sont présentés dans la table 7.3. Nous avons terminé septièmes sur huit participants.

system	dev2010	tst2010	tst2011	tst2012	submission
Baseline (2TM)	27.41	30.28	x	x	
Baseline+GIGAword (3TM)	27.84	30.80	36.88	37.58	**primary**
+DD-google	28.69	**32.01**	39.09	39.36	**contrastive**
+conf	27.84	30.80	x	x	
+DD-google+conf	**28.77**	31.87	x	x	
+DD-ref	32.84	37.26	x	x	oracle
online-google	26.90	33.77	40.16	x	

TABLE 7.3 – Résultats sur la tâche de traduction de la parole anglais-français à la campagne d'évaluation IWSLT 2012 [Besacier et al., 2012].

WMT 2013

Lors de cette campagne, nous avons participé à la tâche d'évaluation de mesures de confiance pour la TA sur la paire de langue espagnol→anglais. Nous prédisons la qualité au niveau du mot, en déterminant s'il est "bon" ou "mauvais" (variante binaire), ou est "bon", ou devrait être "substitué" ou "supprimé" (variante multi-classe). Les résultats sont présentés dans la table 7.4. Nous avons terminé premiers sur cinq participants.

System	Pr	Rc	F	Acc
BOOST_BIN	0.777882	0.884325	0.827696	0.737702
FS_BIN	0.788483	0.864418	0.824706	0.738213
FS_MULT	-	-	-	0.720710
ALL_MULT	-	-	-	0.719177

TABLE 7.4 – Résultats sur la tâche de classification "bon" ou "mauvais" pour chaque mot issu d'un système de TA dans la campagne d'évaluation de WMT13 [Luong et al., 2013b].

WMT 2014

Par rapport à WMT 2013, la tâche est différente en raison d'un système de TA "boîte noire" et de ressources supplémentaires inexistantes. Nous avons fait des efforts pour maintenir le maximum de caractéristiques utilisées en 2013 (en particulier celles du côté source) et proposé de nouvelles relatives aux spécificités du corpus de cette année. Nos résultats ne sont pas en mesure de battre ceux de WMT13, mais restent prometteurs avec ces contraintes. Les

résultats sont présentés dans la table 7.5. Nous avons terminé troisièmes sur cinq participants.

System	F("OK") (%)	Average F(%)
FS(bin) (primary)	74.0961	0.444735
FS(L1)	73.9856	0.317814
FS(mult)	76.6645	0.204953

TABLE 7.5 – Résultats sur la tâche de classification pour la campagne WMT14 [Luong et al., 2014b] "bon" ou "mauvais" pour chaque mot issu d'un système de TA qui est cette fois considéré comme une boîte noire. bin : classification binaire, mult : classification à 3 classes comme à WMT13, L1 : la bonne classe est maintenue intacte, tandis que la mauvaise est divisée en sous-catégories.

IWSLT 2019

Pour cette campagne nous avons participé à la tâche de traduction de textes anglais→tchèque. Nous y avons présenté un système de TA neuronal basé sur l'architecture Transformer. La principale contribution est l'étude de l'impact d'un modèle de langue pré-entraîné en entrée. Nous avons comparé les performances du modèle BERT [Devlin et al., 2019] fourni par les auteurs originaux et d'un modèle BERT contraint que nous avons entraîné sur les données autorisées. Les résultats sont présentés dans la table 7.6. Nous avons terminé cinquièmes sur six participants.

Training data	Input LM	Embeddings size	BLEU (Dev)	BLEU (Test)	TER	BEER	Charac-TER	BLEU (ci)	TER (ci)
MuST-C	None	512	20.4	19.13	61.55	51.64	52.23	19.77	60.54
MuST-C	$BERT_{constr}$	512	20.5	20.02	60.64	51.78	51.26	20.68	59.53
MuST-C	$BERT_{extern}$	512	21.9	21.07	59.19	52.74	49.97	21.78	58.15
MuST-C + News	None	512	22.8	22.26	58.68	53.84	48.29	22.93	57.63
MuST-C + News	$BERT_{constr}$	512	20.4	20.09	60.90	51.91	50.66	20.77	59.79
MuST-C + News	$BERT_{extern}$	512	23.7	23.53	57.55	54.36	47.30	24.27	56.41
MuST-C + News	None	1024	23.1	**22.72**	58.51	53.94	48.27	23.47	57.41
MuST-C + News	$BERT_{constr}$	1024	21.6	21.35	59.68	52.83	49.60	22.05	58.52
MuST-C + News	$BERT_{extern}$	1024	23.9	**23.69**	57.14	54.58	47.06	24.41	56.02

TABLE 7.6 – Résultats sur la tâche de TA anglais→tchèque en utilisant un BERT original (extern) et un BERT appris sur les données autorisées (constr) à IWSLT19 [Vial et al., 2019d].

IWSLT 2020

Lors de cette campagne nous avons participé à deux tâches : traduction de la parole *offline* avec un système de bout-en-bout pour la paire de langue anglais→allemand et traduction de la parole *online* en cascadant un système RAP entraîné à l'aide de *Kaldi* [Povey et al., 2011b] et un système de TA en ligne avec un fonctionnement *wait-k* [Dalvi et al., 2018, Elbayad et al., 2020a].

Les résultats pour le système *offline* sont présentés dans la table 7.7 et ceux du système *online* sont présentés dans la figure 7.2 où l'on voit évoluer le score BLEU en fonction de la latence en terme de nombre mots attendus pour traduire. Nous avons fini cinquièmes sur six pour la première tâche et premiers sur quatre pour la seconde.

2020.contrastive1	18.47	71.85	48.92	55.83	19.46	69.88
2020.contrastive2	19.31	69.30	49.55	52.68	20.36	67.14
2020.contrastive3	20.51	64.88	50.19	53.06	21.5	62.99
2020.contrastive4	15.48	83.45	46.68	57.56	16.42	81.33
2020.contrastive5	16.5	75.15	47.23	57.90	17.42	73.22
2020.primary	**22.12**	**63.87**	**51.20**	**51.46**	**23.25**	**61.85**

TABLE 7.7 – Système *offline* pour la traduction automatique de la parole à IWSLT20 [Elbayad et al., 2020b].

7.2 Réalisations industrielles

Depuis mon arrivée à GETALP j'ai été amené à collaborer avec des industriels pour développer des systèmes dédiés. Je présente ici les principales contributions.

TRIDAN - coordinateur

TRIDAN (Traduction et Reconnaissance d'Images de Documents Arabes Numérisés) vise à optimiser une chaîne de traitement permettant à un utilisateur de faire des requêtes de recherche d'information dans sa langue sur des documents numérisés contenant des informations dans une langue qu'il connaît peu ou pas du tout. Ce projet se déroulant en 2017 est un **partenariat entreprises/LIG**. Il était porté par la société A2IA et ses partenaires sont le LIG, Airbus D&S et l'entreprise Cassidian. Dans le cadre de ce projet nous avons développé un système de traduction automatique de l'arabe [Bouzidi et al., 2017] à partir de documents océrisés.

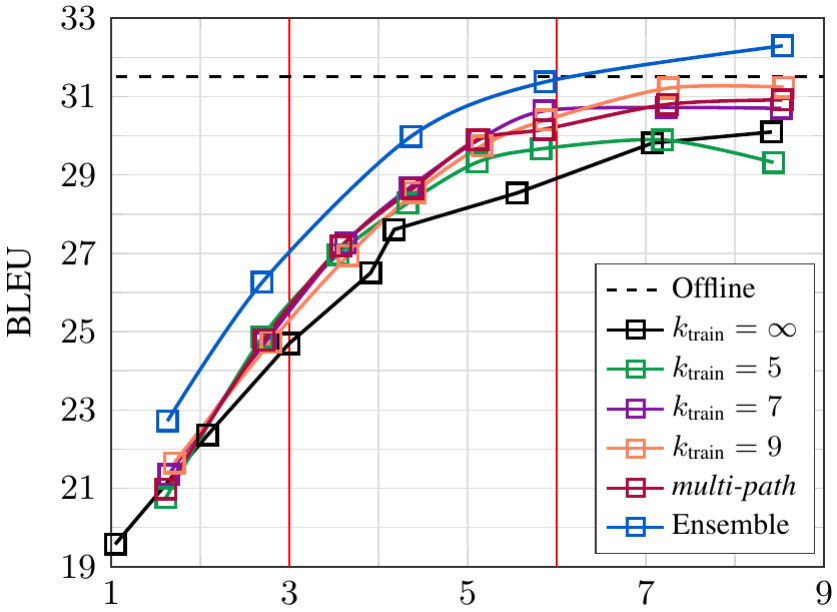


FIGURE 7.2 – Système *online* pour la traduction automatique de la parole avec BLEU en fonction du nombre de mots attendus avant de traduire à IWSLT20 [Elbayad et al., 2020b].

PORTAL - coordinateur

Il s'agit d'un petit projet visant à fournir un système de traduction neuronal arabe/français pour Airbus D&S. Le système a été réalisé par Loïc Vial, en s'appuyant sur une partie des travaux réalisés au sein du projet TRIDAN.

RAP pour les urgences de Genève

Dans le cadre d'un projet [Grijalba Ormaechea et al., 2021] avec l'université de Genève, nous avons développé un système de reconnaissance automatique de la parole Française destiné à être cascadé avec un système de traduction automatique spécialisé pour le diagnostic. La particularité de ce système est qu'il est contraint par une grammaire construite par le laboratoire de traductologie de l'Université de Genève et fonctionne de manière *online*. Ce système vise à remplacer l'utilisation de solutions propriétaires (Nuance) : il est actuellement en phase de validation et présente des résultats de transcription supérieurs à ceux de Nuance. Le travail autour de ce système sera détaillé dans le chapitre 9.

RAP destiné au sous-titrage pour personnes malentendantes

Dans le cadre d'un projet [Evain et al., 2020b] avec l'université de Lyon 2, nous avons développé un système visant à produire des sous-titres pour des personnes malentendantes. Ce système est destiné à être utilisé au sein de l'Université afin que les enseignants puissent transmettre plus naturellement le contenu de leurs cours. Ce système permet d'adapter facilement les modèles de langue à partir du contenu des cours, en s'appuyant notamment sur des données textuelles extraites de Wikipédia. Le schéma global du système est présenté dans la figure 7.3. Ce système fonctionne de manière *online*.

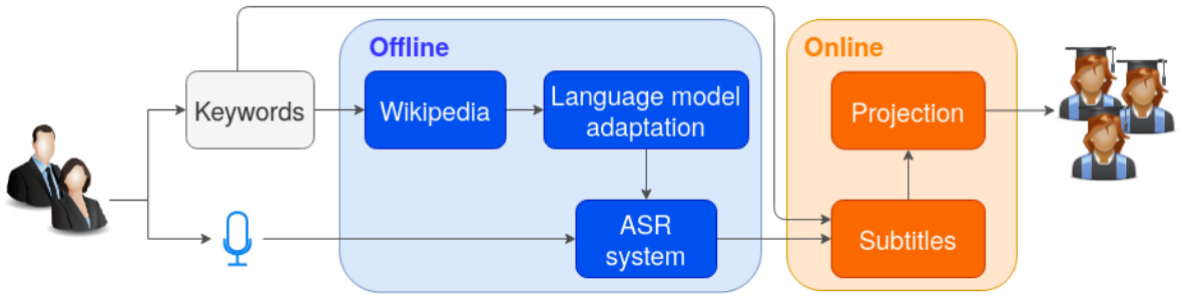


FIGURE 7.3 – Système RAP basé sur *Kaldi* et destiné à fournir un sous-titrage dans le cadre de cours donnés à des personnes malentendantes. Figure extraite de [Evain et al., 2020b].

7.3 Conclusion

Dans ce chapitre nous avons présenté un autre aspect de la recherche en informatique, à savoir les campagnes d'évaluation et les réalisations industrielles. Elles permettent d'apporter de la visibilité à l'équipe et de maintenir à jour les outils que nous développons ; la valorisation de ces activités au niveau de la recherche n'est pas toujours évident, bien qu'elles soient très chronophages. Le chapitre suivant présentera les valorisations s'articulant autour des corpus et outils libres que nous diffusons.

Chapitre 8

Corpus et outils partagés avec la communauté

Encadrements : Solène Evain, Ngoc Tien Le, Ngoc Quang Luong, Lucia Ormaechea, Pauline Trial, Céline Vaschalde, Loïc Vial

Sélection de publications relatives à ce chapitre :

S. Evain, H. Nguyen, H. Le, M. Zanon Boito, S. Mdhaffar, S. Alisamir, Z. Tong, N. Tomashenko, M. Dinarelli, T. Parcollet, A. Allauzen, Y. Estève, **B. Lecouteux**, F. Portet, S. Rossato, F. Ringeval, D. Schwab, L. Besacier. LeBenchmark : A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech. Interspeech 2021.

S. Evain, **B. Lecouteux**, F. Portet, I. Estève, M. Fabre. Towards Automatic Captioning of University Lectures for French Students who are Deaf (papier court). ACM SIGACCESS 2020.

D. Schwab, P. Trial, C. Vaschalde, L. Vial, E. Esperança-Rodier, **B. Lecouteux**. Providing semantic knowledge to a set of pictograms for people with disabilities : a set of links between WordNet and Arasaac : Arasaac-WN. LREC 2020.

H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, **B. Lecouteux**, A. Allauzen, B. Crabbé, L. Besacier, D. Schwab. . FlauBERT : Unsupervised Language Model Pre-training for French. LREC 2020.

L. Vial, **B. Lecouteux**, D. Schwab. UFSAC : Unification of Sense Annotated Corpora and Tools. LREC 2018.

M. Vacher, S. Bouakaz, ME. Bobillier-Chaumon, F. Aman, RA. Khan, S. Bekkadja, F. Portet, E. Guillou, S. Rossato, **B. Lecouteux**. The CIRDO Corpus : Comprehensive Audio/Video Database of Domestic Falls of Elderly People. LREC 2016.

NT. Le, C. Servan, **B. Lecouteux**, L. Besacier. Better Evaluation of ASR in Speech Translation Context Using Word Embeddings. InterSpeech 2016.

C. Servan, NT. Le, NQ. Luong, **B. Lecouteux**, L. Besacier. An Open Source Toolkit for Word-level Confidence Estimation in Machine Translation. IWSLT 2015.

Ce chapitre aborde quelques diffusions de corpus et d'outils auxquels j'ai contribué, les liens vers ces travaux sont disponibles à partir de la page de l'équipe http://lig-getalp.imag.fr/demos/. Ces corpus et outils sont importants dans le domaine du TALN. Leur mise en place représente un travail conséquent. La politique de l'équipe est de diffuser ces corpus afin d'assurer la reproductibilité de la recherche.

8.1 Arasaac-Wordnet

Arasaac-Wordnet [Schwab et al., 2019, Schwab et al., 2020] est une ressource reliant WordNet et Arasaac (en l'occurrence une grande base de données, librement accessible, de pictogrammes). Les pictogrammes représentent un outil très utilisé par les personnes souffrant de troubles cognitifs ou de la communication. Malheureusement, leur exploitation est manuelle via des classeurs, alors que les soignants et les familles souhaiteraient avoir accès à des outils plus automatisés, comme par exemple la génération de pictogrammes par la parole. Pour ce faire, nous proposons l'usage automatique de pictogrammes dans les applications NLP, à partir d'une base de données reliant lesdits pictogrammes à des connaissances sémantiques. Cette ressource sera importante pour la création d'applications d'aide aux personnes souffrant de troubles cognitifs, telles que text→picto, speech→picto ou picto→speech. À l'heure actuelle, cette ressource contient environ 800 pictogrammes associés à leurs *synsets WordNet*. Elle est accessible via une collection numérique (basée sur Omeka) et via une base de données SQL.

8.2 Corpus pour Sweet-home

Dès mon arrivée dans l'équipe GETALP, j'ai commencé à travailler sur la mise en place d'un protocole et d'un corpus [Vacher et al., 2014a, Vacher et al., 2016a] visant à tester nos systèmes de Reconnaissance de la parole dans un habitat intelligent et dans le cadre du projet Sweet-Home. J'ai notamment participé à la création du corpus "Home Automation Speech" qui a été enregistré pour développer une reconnaissance automatique robuste des commandes vocales dans une maison intelligente dans des conditions distantes (les microphones étant au plafond). Huit canaux audio ont été enregistrés pour acquérir un corpus vocal représentatif composé d'énoncés de commandes domotiques et d'appels de détresse, mais aussi de phrases familières. Le dernier microphone a enregistré spécifiquement la source de bruit pour les expériences d'annulation du bruit. Afin d'obtenir des conditions plus réalistes, deux types de bruits de fond ont été pris en compte pendant que l'utilisateur parlait : une radio d'informations et une musique classique. Le corpus est composé, pour chaque locuteur, d'un texte de 285 mots pour l'adaptation acoustique (36 minutes pour 351 phrases au total pour 23 locuteurs), et de 240 phrases courtes (2 heures et 30 minutes par canal au total 23 locuteurs) avec un total de 5520 phrases au total.

8.3 FLUE

Dans le cadre du projet Flaubert, il nous est apparu indispensable de fournir à la communauté scientifique un ensemble de tâches représentatives du traitement automatique des langues du français. Nous fournissons les corpus (que nous avons soit créés, soit rassemblés ici) ainsi que le code permettant de réaliser les tests. Cette *baseline* nous a permis de confronter nos modèles Flaubert à ceux de l'état-de-l'art et devrait constituer une référence pour le français dans les années qui viennent. https://github.com/getalp/Flaubert/tree/master/flue. Parmi les *baselines* mises à disposition j'ai particulièrement participé à celle portant sur la désambiguïsation lexicale, dans le cadre de la thèse de Loïc Vial [Vial, 2020].

8.4 UFSAC : Corpus pour la désambiguïsation lexicale

Durant la thèse de Loïc Vial portant sur la désambiguïsation lexicale, ses travaux [Vial et al., 2017a, Vial et al., 2017b, Vial et al., 2018a, Vial et al., 2019a, Vial et al., 2019b, Vial et al., 2019c, Vial et al., 2019d] ont mis en évidence des problèmes de normalisation des corpus et une difficulté à les exploiter. Nous avons ainsi proposé [Vial et al., 2017c, Vial et al., 2018b] de mettre à disposition des outils pour travailler aisément avec les corpus existants dédiés à la désambiguïsation lexicale. En effet, il existe aujourd'hui une douzaine de corpus anglais annotés en sens, dans des formats différents les uns des autres. Nous avons donc proposé UFSAC : un format de corpus pouvant être utilisé pour l'entraînement ou le test d'un système de désambiguïsation. Nous fournissons à la communauté l'ensemble des corpus anglais annotés de sens que nous connaissons au format UFSAC, et utilisant la dernière version de *WordNet*. Nous fournissons également le code source permettant de générer les corpus à partir de leurs données originales et une API Java permettant de manipuler les corpus au format UFSAC `https://github.com/getalp/UFSAC`.

8.5 Corpus pour la traduction automatique de la parole

Dans le cadre de la thèse de Ngoc Tien Le [Le, 2018] nous avons travaillé sur les mesures de confiance pour la traduction automatique de la parole. Il n'existait pas de corpus existant : une partie de sa thèse a porté sur la création d'un corpus afin d'avoir les correspondances orales d'un corpus textuel français déjà en notre possession. Nous avons réalisé des enregistrements avec des volontaires natifs. À chaque énoncé est associé un quintuplet : la sortie du SRAP (f_{hyp}), la transcription verbatim (f_{ref}), la sortie de la traduction du texte source vers l'anglais ($e_{hyp_{mt}}$), la sortie de la traduction de la parole ($e_{hyp_{slt}}$) et la post-édition de la traduction (e_{ref}). Ce corpus est disponible sur `https://github.com/besacier/WCE-SLT-LIG/`. La durée totale du corpus de parole obtenu est de 16h52. Ce corpus et sa conception sont plus détaillés dans la section 6.2 et dans [Le et al., 2016a, Le et al., 2018].

8.6 Toolkit pour générer des mesures de confiance

Dans [Servan et al., 2015], nous présentons un *toolkit* open-source dédié aux mesures de confiance. Il permet de combiner à la fois des caractéristiques externes et des caractéristiques internes. Toutes ces caractéristiques extraites sont combinées via un classificateur à base de CRF pour estimer la confiance d'un mot. Les expériences réalisées avec cet outil sont état-de-l'art et reproductibles. Elles s'articulent autour de deux ensembles de données correspondant à deux paires de langues (français→anglais et anglais→espagnol). Le *toolkit* a été rendu flexible pour le rendre indépendant de la langue, pour autant que l'utilisateur puisse fournir des modèles pour de nouvelles langues. Nous y avons intégré un module permettant de sélectionner automatiquement les caractéristiques en utilisant l'algorithme SFS (*Sequential Forward Selection*). Ce *toolkit* est disponible sur `https://github.com/besacier/WCE-LIG` et son fonctionnement global est présenté dans la figure 8.1.

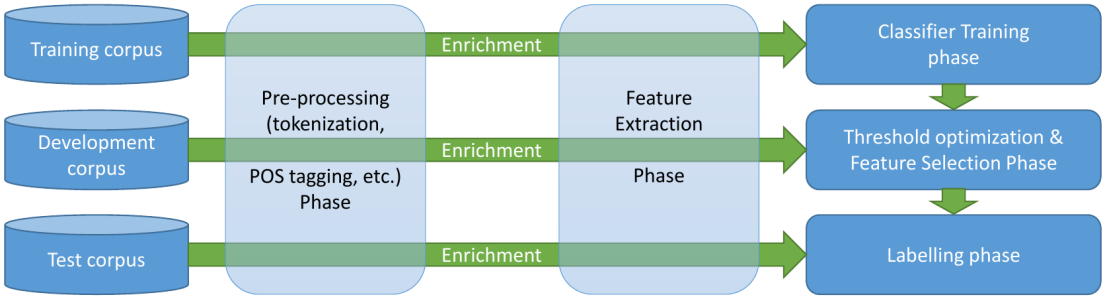


FIGURE 8.1 – Schéma du fonctionnement du *toolkit* que nous avons mis à disposition de la communauté [Servan et al., 2015].

8.7 Mesure sémantique de la qualité de traduction automatique

Dans le cadre du projet KEHATH nous avons proposé une nouvelle métrique [Le et al., 2016b] se basant sur des plongements de mots : nous proposons une extension du WER pénalisant les erreurs de substitution en fonction de leur contexte (à base de plongements de mots). Nos expériences, axées sur une tâche de traduction de la parole français→anglais, prouvent que la nouvelle métrique proposée est mieux corrélée aux performances. Ces mêmes expériences ont

également montré la capacité de la métrique à sérier de meilleures hypothèses, dans l'optique de la traduction, parmi les*N-best* du SRAP. Ce travail est détaillé dans la section 6.1. En ce qui concerne la reproductibilité, l'outil est mis à disposition sur https://github.com/cservan/tercpp-embeddings.

8.8 Lebenchmark

Lebenchmark est issu d'un projet qui rassemble 18 chercheurs et 3 laboratoires d'informatique [Evain et al., 2021b] autour des tâches qui utilisent la reconnaissance automatique de la parole. Depuis quelques années, l'apprentissage auto-supervisé sur de grandes quantités de données non annotées s'est démocratisé dans les domaines de l'image et du TALN. En 2019, ces approches auto-supervisées se sont étendues à la parole. Elles sont de plus en plus fréquentes, mais les protocoles de comparaison présentent des faiblesses : les paramètres des modèles ne sont pas facilement reproductibles, les corpus d'évaluation ne sont pas toujours disponibles, les métriques d'évaluation ou les normalisations diffèrent, etc. De plus, la plupart des corpus disponibles le sont pour l'anglais.

Ces multiples obstacles rencontrés mettent en exergue la difficulté de comparer objectivement différentes approches et d'évaluer leur impact sur tel ou tel système de reconnaissance automatique de la parole. Partant de ce constat nous avons proposé un cadre reproductible permettant d'évaluer les approches auto-supervisées et incluant le maximum de tâches liées à la parole. Nous fournissons pour chaque tâche : les modèles, les corpus et les outils d'évaluation. Par ailleurs, ce travail est dédié à la langue française.

Mes contributions portent sur une partie de la collecte du corpus et la partie *benchmark* relative aux modèles SRAP hybrides.

Données

LeBenchmark est la première étude proposant un ensemble de données aussi vaste et diversifié pour l'entraînement d'un système auto-supervisé pour le français. Le corpus comprend 2933h de parole, dont 1115h de parole lue, 1626h de parole diffusée, 124h de parole spontanée, 38h de dialogues téléphoniques joués et 29h de parole émotionnelle jouée. En ce qui concerne le genre, nous avons recueilli 1824 heures de discours issus de locuteurs, 1034 heures de locutrices et 74 heures pour lesquelles nous n'avons pas l'information.

Modèles auto-supervisés pour le français

LeBenchmark fournit quatre modèles de type Wav2Vec2.0 [Baevski et al., 2020] pré-entraînés sur les données décrites ci-dessus. Nous proposons comme [Baevski et al., 2020], deux architectures différentes : *large* et *base* qui sont couplées à nos modèles. Ces modèles sont partagés avec la communauté via HuggingFace : https://huggingface.co/LeBenchmark avec pour objectif une recherche reproductible.

Nous proposons également un set de *benchmarks* pour les tâches suivantes :

- La reconnaissance automatique de la parole fonctionnant avec des systèmes hybrides (HMM-DNN).
- La reconnaissance automatique de la parole de bout-en-bout.
- Les systèmes de compréhension de la parole.
- Les systèmes de traduction automatique de la parole.
- Les systèmes de reconnaissance des émotions sur le Français.

Le lecteur trouvera des résultats exhaustifs sur les modèles créés dans [Evain et al., 2021b].

8.9 Conclusion

Dans ce chapitre nous avons présenté quelques outils et corpus auxquels j'ai contribué. Ils sont indispensables dans le domaine du TALN et représentent souvent une charge conséquente de travail pour être mis à disposition de la communauté. La reproductibilité de la recherche est devenue de plus en plus prépondérante ces dernières années et le partage des données nécessaires à obtenir les résultats présentés dans nos articles répond en partie à cette problématique.

Chapitre 9

Travaux en cours préliminaires : parole, pictogrammes et urgences

Encadrements : Lucia Ormaechea, Pauline Trial, Céline Vaschalde, Loïc Vial

Sélection de publications relatives à ce chapitre :

L. Ormachea Grijalba, P. Bouillon, J. Gerlach, **B. Lecouteux**, D. Schwab, H. Spechbach. Reconnaissance vocale du discours spontané pour le domaine médical. Conférence sur les Technologies du Langage Humain. TLH 2021.

D. Schwab, P. Trial, C. Vaschalde, L. Vial, E. Esperança-Rodier, **B. Lecouteux**. Providing semantic knowledge to a set of pictograms for people with disabilities : a set of links between WordNet and Arasaac : Arasaac-WN. LREC 2020.

C. Vaschalde, P. Trial, E. Esperança-Rodier, D. Schwab, **B. Lecouteux**. Automatic pictogram generation from speech to help the implementation of a mediated communication. Conference on Barrier-free Communication 2018.

Avec l'arrivée du *deep-learning*, la communauté parole a glissé vers le *machine-learning*. La reconnaissance automatique de la parole a changé de paradigme en devenant une tâche de classification agnostique où les approches sciences humaines et sociales (SHS) ont été un peu écartées. En 2014 j'ai commencé à travailler plus étroitement avec Didier Schwab qui commençait à développer ses recherches autour de la communication alternative et augmentée (CAA). Ce type de recherche m'a de suite intéressé et j'ai été conquis par la pluridisciplinarité du domaine (linguistes, phonéticiens, orthophonistes, psychologues, informaticiens etc.). J'y ai trouvé la possibilité de me rapprocher à nouveau des SHS pour lesquelles j'ai toujours éprouvé un réel intérêt, tout en exploitant les nouveaux usages rendus possibles par les approches issues du *machine-learning*. Mes thématiques de recherche se sont alors, petit à petit, orientées en direction de la traduction automatique de la parole vers des pictogrammes.

9.1 Traduction automatique de la parole et pictogrammes

Lorsque l'utilisation de la parole ou de la langue des signes pour communiquer est impossible en raison d'aphasie, de dysarthrie et de troubles physiques, les personnes ne sont pas en mesure d'exprimer leurs sentiments ou leurs besoins et ne peuvent créer aucun lien social, ni communiquer.

L'utilisation de méthodes de communication alternative et améliorée (CAA) est un moyen d'aider ces personnes. Ces méthodes remplacent ou soutiennent les capacités de parole d'un locuteur. Elles utilisent souvent l'encodage visuel des informations, notamment via des pictogrammes qui sont plus iconiques que les mots en raison de leur ressemblance avec le référent [Duboisdindien, 2014].

Le pictogramme peut être défini, en CAA, comme un signe graphique dont le signifiant présente une similitude plus ou moins forte avec le signifié, contrairement aux signes linguistiques graphiques ou phoniques dont la forme du stimulus est souvent indépendante de celle du référent. Il permet une représentation plus iconique de l'information et est donc plus facilement interprétable. Néanmoins, la façon dont les personnes interprètent un pictogramme peut être extrêmement variable en raison de l'ensemble des pictogrammes utilisés, du contexte culturel et de la signification du pictogramme (les pictogrammes grammaticaux sont plus complexes à comprendre car ils sont moins iconiques).

Les pictogrammes, grâce à leur iconicité, peuvent aider les gens à communiquer dans un pays étranger lorsqu'ils ne parlent pas la langue locale et

ne partagent aucun bagage linguistique commun avec les habitants. Comme [Mihalcea, 2008] l'ont montré, la traduction de pictogrammes peut aider les personnes qui ne partagent pas la même langue à communiquer.

9.2 Travaux préliminaires autour de la parole et des pictogrammes

Afin de construire des phrases à l'aide de pictogrammes et d'augmenter la taille du vocabulaire du locuteur, il est nécessaire de disposer d'un apport riche en pictogrammes de phrases provenant de la famille [Beukelman et al., 2017]. Des classeurs de communication (sur support papier ou électronique) sont souvent utilisés pour encoder ces phrases. Trouver le pictogramme requis dans un classeur de communication est une tâche difficile. La famille doit apprendre à utiliser l'outil de communication et doit consacrer du temps à la recherche des pictogrammes. En raison de cette étape de recherche, l'interaction n'est pas spontanée et nuit à la qualité de la communication.

Un outil de génération automatique de pictogrammes fonctionnant avec le langage courant est un bon moyen de résoudre ce problème. Un tel outil permettrait aux personnes proches des personnes en situation de handicap de parler avec des pictogrammes sans avoir à apprendre comment les coder et sans perdre de temps à les trouver dans un classeur de communication. Cela permettrait un meilleur accès aux écoles pour les utilisateurs de moyens de CAA.

La génération de pictogrammes permet donc de surmonter la barrière de la langue entre les personnes, et peut permettre aux personnes de rejoindre une école ou un cours de formation plus facilement qu'auparavant. Il s'agit d'un besoin observé aussi bien dans les structures d'accueil des personnes en situation de handicap que chez les proches : elles ne peuvent pas communiquer avec leur environnement de manière traditionnelle avec leur voix, et parfois même pas avec des gestes.

À l'aune des travaux de Céline Vaschalde [Vaschalde et al., 2018a, Vaschalde et al., 2018b] nous proposons d'utiliser un module de reconnaissance automatique de la parole permettant de travailler directement avec la voix. Il prend un signal vocal et le transforme en une transcription orthographique. Le modèle de RAP est basé sur un modèle hybride HMM-DNN, développé par [Elloumi et al., 2018c] avec la boîte à outils *Kaldi* [Povey et al., 2011b].

Nous présentons ainsi des premiers résultats texte→pictogramme. Le prototype obtient un score BLEU de 26,65 lorsque tous les mots sont traduits et de

19,91 lorsque le texte est simplifié (certains mots grammaticaux sont supprimés). Cette évaluation met en évidence les difficultés rencontrées dans la tâche de simplification du texte. En effet, il est particulièrement difficile d'identifier les mots grammaticaux qui peuvent être supprimés de ceux dont la suppression peut changer significativement le sens du message. Lorsque nous avons construit notre corpus d'évaluation simplifié, de nombreux cas complexes de suppression d'adverbes nous ont posé question et certains ont dû être conservés afin de ne pas trop modifier le sens du texte. Ces travaux préliminaires sont à l'origine du projet ANR PROPICTO présenté dans la section 10.1.

9.2.1 Un système de RAP pour les urgences de Genève

Contexte

Les travaux de Céline Vaschalde ont permis de publier un article dans la conférence *Barrier free conference* [Vaschalde et al., 2018b]. Lors de cette conférence nous avons rencontré l'équipe TIM de Pierrette Bouillon qui commençait à travailler sur la traduction vers des pictogrammes, mais dédiée au médical. Nous avons entamé une collaboration autour de ce sujet, en commençant par les aspects liés à la reconnaissance automatique de la parole. Nous avons ainsi co-encadré le stage de M2 recherche de Lucia Ormaechea visant à travailler sur la reconnaissance automatique de la parole pour le milieu médical et plus précisément un service d'urgences. Ses travaux sont détaillés dans la section qui suit.

Barrières linguistiques et urgences au CHU de Genève

Dans les services d'urgence, la barrière linguistique peut constituer un problème car la difficulté ou l'impossibilité d'une bonne interaction entre une personne soignée et le médecin peut avoir des répercussions gravissimes : [Hacker et al., 2015]. C'est précisément pour cette raison que des protocoles fiables de traduction et d'interprétation médicale sont importants.

C'est dans cette optique qu'est né le projet BabelDr [Bouillon et al., 2017] porté par Pierrette Bouillon, afin de garantir une assistance efficace dans des contextes d'urgence médicale multilingue et d'éliminer autant que possible les barrières linguistiques entre médecin et patient. La voix du locuteur est reconnue à l'aide d'un système de reconnaissance automatique de la parole actuellement fourni par *Nuance*. La transcription est ensuite liée à une forme canonique, qui sert de pivot pour la traduction et de rétro-traduction pour

vérifier la reconnaissance. Une fois validée elle est considérée de manière sûre comme l'entrée à traduire et à lire à haute voix dans la langue cible.

BabelDr propose actuellement des traductions dans des langues telles que l'arabe, l'espagnol ou le farsi, mais aussi dans des langues plus rares comme le tigrinya ou l'albanais. Afin d'améliorer son accessibilité, BabelDr travaille à l'inclusion de la langue des signes des sourds de Suisse romande et envisage d'évoluer vers un système qui traduirait également les énoncés en pictogrammes : [Vaschalde et al., 2018b, Norré et al., 2020, Schwab et al., 2020]. Ce système serait utile non seulement comme moyen de communication patient-médecin pour les locuteurs allophones (à savoir les non-francophones), mais aussi pour les personnes souffrant d'un handicap cognitif.

Un SRAP contraint par une grammaire experte

Lors du M2 recherche de Lucia Ormaechea, nous avons travaillé sur la mise en place d'un SRAP robuste destiné à remplacer les outils fournis par *Nuance* : l'objectif était d'avoir un système contraint par une grammaire exprimant les différentes formes de phrases diagnostics.

Afin de construire ce système SRAP à des fins médicales et cascadé avec BabelDr, nous avons choisi d'utiliser *Kaldi* [Povey et al., 2011b] qui a l'avantage d'être basé sur des transducteurs d'états finis (FST). Les travaux que nous avons proposés se sont appuyés sur ceux de [Horndasch et al., 2016].

Le laboratoire de traductologie de Genève a conçu en partenariat avec le CHU de Genève des grammaires modélisant les questions aidant au diagnostic. L'objectif de notre travail a été d'intégrer ces grammaires qui sont représentées dans un formalisme Regulus Lite [Rayner et al., 2016] au sein de *Kaldi*. Nous avons distingué un modèle de langue principal et des modèles de sous-langue. Le modèle principal est composé d'énoncés modélisant le discours médical en considérant différentes paraphrases, et le sous-langage correspond à des classes : *TrLex* pour les paradigmes lexicaux et *TrPhrase* pour les modèles phrastiques. Ces deux classes sont identifiées par des symboles non terminaux et peuvent être intégrées dans le modèle de langue principal. Un exemple du format du fichier source est le suivant :

- Utterance : Utterance $avez_vous (mal | douleur) (quelque part | à un endroit) EndUtterance
- Phrasal patterns : TrPhrase $avez_vous (avez-vous | vous avez) EndTrPhrase

Afin de générer la grammaire sous forme de FST utilisable dans *Kaldi* nous devons suivre les étapes suivantes :

1. Convertir chaque motif en un FST
2. Tous les FST appartenant au modèle de langue principal sont unifiés en un seul FST, tandis que pour les modèles de sous-langues, une union de chaque classe de mots est effectuée. Si l'on reprend l'exemple précédent, le FST résultat est présenté dans la figure 9.1.

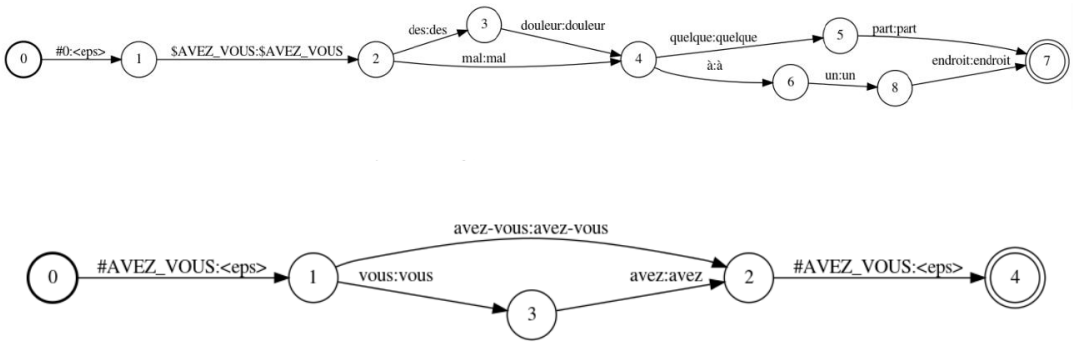


FIGURE 9.1 – Equivalent FST de la phrase donnée en exemple au format Regulus Lite.

3. Lors de l'intégration des classes de mots dans la grammaire principale, les symboles non-terminaux existants sont remplacés récursivement dans le modèle de langue principal.

Il résulte de ces étapes une grammaire FST contenant la grammaire principale, ce qui donne pour notre exemple le FST présenté dans la figure 9.2 :

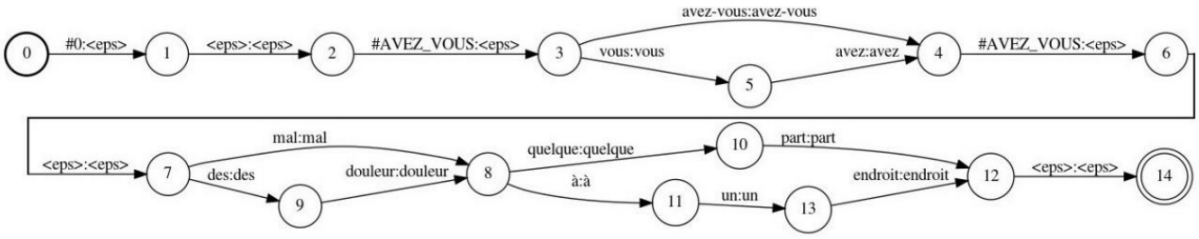


FIGURE 9.2 – Grammaire compilée avec les états terminaux.

Tous les outils pour reproduire les expériences sont disponibles sur https://github.com/lormaechea/Grammar-Tools-ASR

Résultats

Notre système de reconnaissance vocale est basé sur un modèle hybride HMM-DNN construit avec *Kaldi* [Povey et al., 2011b]. Le modèle acoustique est

de type TDNN [Povey et al., 2018] avec des i-vectors pour l'adaptation au locuteur [Dehak et al., 2010]. Les données acoustiques d'entraînement comprennent 500 heures de parole française : ESTER [Galliano et al., 2006], COMMON-VOICE [Ardila et al., 2019] et ESLO [Serpollet et al., 2007]. Il est à noter que le système de reconnaissance fonctionne en temps réel et produit des hypothèses à la volée.

		WER(%)		**SemER(%)**[1]	
Corpus	***Phrases***	**Nuance**	**Kaldi**	**Nuance**	**Kaldi**
Dev	**2864**	20,99	14,15	17,59	21,11
Test	**2708**	22,93	14,37	35,52	14,73

TABLE 9.1 – Résultats de notre système comparé à celui de *Nuance* : le dev et le test représentent chacun 1 heure de signal audio. SemER indique si le système a trouvé la bonne phrase canonique.

La table 9.1 présente les performances de notre système basé sur les grammaires construites par le laboratoire de Genève. Les corpus utilisés ont été enregistrés par le CHU et l'Université de Genève : il s'agit de situations simulées avec des qualités d'enregistrement variées (souvent sur ordinateur ou smartphone). Nous observons que notre système est meilleur que *Nuance* en termes de WER et le semER qui représente les erreurs au niveau de la détection de la phrase diagnostic (mesure basée sur une distance cosine sur des vecteurs de type BERT) s'est dégradé sur le dev mais est divisé par deux sur le test. Cependant, le système de détection de phrase diagnostic a été entraîné sur les données de *Nuance* et non les nôtres (nous n'avions pas la main sur ce module), ce qui biaise sans doute la qualité des résultats. Par ailleurs les conditions de test sont plus difficiles et proches de la réalité par rapport au dev. Actuellement notre système est en cours de validation pour passer en production au sein des urgences du CHU de Genève et sera exploité dans le cadre du projet PROPICTO.

9.2.2 Lier Wordnet avec des pictogrammes

En 2020, nous avons proposé, dans [Schwab et al., 2020], d'établir une relation entre deux ressources très utilisées dans leurs domaines respectifs : d'une part le *Princeton WordNet* et d'autre part le jeu de pictogrammes *Arasaac*. En effet, à l'heure actuelle, il est difficile d'utiliser automatiquement le jeu de pictogrammes *Arasaac* dans un pipeline de traitement du langage naturel (*Natural Language Processing* NLP), car aucune information sémantique associée

n'est disponible. Généralement, les utilisateurs de pictogrammes les choisissent en fonction d'une description textuelle, d'une représentation graphique, d'un manuel d'utilisation ou suite à une formation.

Nous présentons donc une première base de données faisant ce lien à partir de pictogrammes qui sont à la fois libres et utilisés dans les institutions. L'objectif est de favoriser le développement d'outils autour des pictogrammes en souhaitant que cette base s'enrichisse au fil des années.

Les pictogrammes Arasaac sont indépendants de la langue, et la plupart sont étiquetés avec des légendes multilingues (anglais, français, espagnol...). Afin d'assister le processus de liaison d'un pictogramme à *WordNet*, nous avons utilisé un système de désambiguïsation lexicale français, qui donne le sens *WordNet* le plus probable à chaque mot pour une phrase française donnée.

Pour la construction de ce système, nous avons exploité la méthode proposée par [Hadj Salah et al., 2018], qui consiste à traduire et aligner automatiquement des corpus annotés de sens anglais vers le français, afin d'obtenir des corpus annotés de sens français. En effet, les données annotées manuellement sont rares et presque inexistantes dans les langues non anglaises. Nous nous sommes alors appuyés sur le système WSD état-de-l'art [Vial et al., 2019c] conçu durant la thèse de Loïc Vial [Vial, 2020] et qui a été adapté au français.

Une fois le système WSD entraîné, nous avons utilisé ses prédictions pour faciliter le travail de mise en correspondance des pictogrammes avec *WordNet*. Par exemple, si l'annotateur voulait mettre en correspondance un pictogramme pour "prendre une douche", le sens prédit par le système WSD sur le mot "prendre" peut l'aider à adopter une décision.

Actuellement, environ 800 pictogrammes ont été reliés à *WordNet* ce qui donnera accès à une gamme d'outils de traitement automatique du langage pour aider les personnes en situation de handicap. La réalisation de ce type de base est assez fastidieuse en raison des difficultés rencontrées et du protocole d'annotation. À l'origine, nous avions commencé à développer cette base de données pour nos recherches, mais nous avons souhaité la partager avec la communauté afin que les outils et les travaux sur la communication alternative via les pictogrammes puissent être développés rapidement et de manière collaborative.

9.3 Conclusion

Dans ce chapitre nous avons abordé quelques travaux préliminaires autour de la CAA qui se sont plus particulièrement axés sur la traduction automatique de

la parole vers des pictogrammes et les systèmes RAP spécialisés pour le milieu médical. Depuis 2018, nous avons réalisé avec Didier Schwab plusieurs travaux qui ont servi de socle pour monter le projet ANR international PROPICTO. Ainsi nous avons soulevé de nombreux verrous scientifiques (et techniques) qui existent pour réaliser un tel projet. La rencontre avec Pierrette Bouillon a été très enrichissante car nos domaines de recherche sont extrêmement complémentaires. Cette collaboration a abouti avec le projet ANR PROPICTO qui est présenté dans le chapitre suivant.

Chapitre 10

Travaux à venir et conclusion

10.1 Le projet ANR PROPICTO

Suite à nos recherches autour de la CAA et à la rencontre avec Pierrette Bouillon, nous avons monté, avec la Suisse, un projet ANR international dont je suis porteur et coordinateur pour la France. Ce projet permet de lier les travaux de Pierrette Bouillon autour de la traduction de la parole destinée à des personnes allophones dans un cadre médical et nos travaux de traduction de la parole vers des pictogrammes destinée aux personnes en situation de handicap. Le projet PROPICTO (PRojection du langage Oral vers des unités PICTOgraphiques) a démarré dans le courant de l'année 2021.

10.1.1 Résumé du projet [1]

PROPICTO vise à développer un axe de recherche autour de la communication alternative et augmentée, en se focalisant sur la transcription automatique de la parole française sous forme pictographique. Il répond ainsi à de nombreux besoins sociétaux dans le domaine du handicap (communiquer avec des personnes ayant des problèmes cognitifs) et médical (communiquer avec des patients qui n'ont pas la même langue que le praticien). Il prend également en compte les exigences légales adoptées tant en Suisse (loi fédérale sur l'élimination des inégalités frappant les personnes handicapées de 2002, ainsi que la Convention de l'ONU relative aux droits des personnes handicapées, ratifiée par la Suisse en 2014) qu'en France (loi du 2 janvier 2002, renforcée par la loi du 11 février 2005).

PROPICTO relève de nombreux défis de recherche autour du traitement

1. Résumé officiel extrait du projet ANR https://anr.fr/Projet-ANR-20-CE93-0005

automatique de la langue naturelle. La finalité du projet est de proposer des méthodes et des corpus qui permettront de transcrire directement la parole vers une suite de pictogrammes, libres (ARASAAC) ou spécialement créés pour les besoins (médical, familial etc.). Nous partirons d'un domaine spécialisé (urgences médicales) pour l'étendre aux domaines spontanés utilisés dans les instituts ou bien en famille.

Le projet devra intégrer deux contraintes majeures : 1) la faible quantité de données qui est un frein à la mise en œuvre des techniques état de l'art à base d'apprentissage automatique et 2) la nécessité d'évaluer nos méthodes avec des groupes cibles très diversifiés. Il adoptera une approche modulaire qui traitera indépendamment les quatre étapes du projet, à savoir :

- reconnaissance automatique de la parole spontanée,
- analyse syntaxique de l'oral,
- simplification de l'oral vers une norme : Facile À Lire et à Comprendre (FALC) et
- traduction du FALC en pictogrammes.

Chaque étape s'appuiera sur des approches hybrides où des règles linguistiques serviront d'amorce à des systèmes robustes, afin de pallier le manque de données initial. Ces expertises linguistiques passeront, par exemple, par des grammaires expertes, la génération synthétique de corpus et la modélisation syntaxique experte. Cette approche modulaire facilitera également l'évaluation des différentes étapes en fonction des groupes cibles. Les besoins pourront ainsi différer à différents niveaux : langage spécialisé/simple/spontané, syntaxe complexe/simplifiée et pictogrammes spécialisés/génériques/avancés/simplifiés.

PROPICTO mettra à disposition de la communauté scientifique l'ensemble des ressources créées : corpus audio associé à sa traduction en pictogrammes (avec différentes situations écologiques), base de données liant pictogrammes et leur signification sémantique, systèmes de Reconnaissance Automatique de la Parole (RAP) du projet, système de simplification pour le FALC, système de Traduction Automatique (TA) parole/pictogramme et métriques d'évaluation humaine ou automatique de la traduction parole/pictogramme. À l'issue du projet, trois prototypes destinés à des publics cibles différents, seront mis en production : 1) pour les urgences aux "Hôpitaux Universitaires de Genève" (HUG, Suisse), 2) en institution au sein de l'Établissement Coste-Rousse pour Enfants et Adultes Polyhandicapé (EEAP, France) et 3 instituts français médico-éducatifs (IME) de Valence, Orange, Meylan dans le cadre familial/quotidien auprès de volontaires de l'Association Française du Syndrome de Rett (AFSR).

ces prototypes seront ainsi testés en conditions réelles et évalués avec des méthodes humaines et automatiques.

10.1.2 État de la recherche dans le domaine

Le projet vise à fournir une aide à la communication médiée par la machine, pour la langue française. Nous nous intéressons spécifiquement aux troubles du langage physiques et/ou cognitifs, où une interaction vocale directe n'est pas envisageable.

Pour permettre aux individus en situation de handicap langagier de communiquer, plusieurs méthodes de Communication Alternative et Augmentée (CAA) [Nègre and Superieur, 2017] peuvent être utilisées. Celles-ci sont *alternatives* lorsqu'elles remplacent totalement les moyens d'expression orale et *améliorées* lorsqu'elles permettent de compléter les capacités de communication existantes, pouvant aider dans certains cas à l'émergence de la graphie ou de l'oralisation [Beukelman et al., 2017].

Le cœur du projet s'articule autour de l'assistance à la communication orale. Les personnes en situation de handicap ne peuvent pas communiquer avec leur environnement de manière classique par l'intermédiaire de la voix, ni même parfois avec une gestuelle. Certes d'un individu à l'autre les capacités cognitives diffèrent, mais la mise en place d'un système de communication même rudimentaire est presque toujours possible ; la communication via des pictogrammes est ainsi de plus en plus utilisée dans les institutions et commence à se normaliser [Bandeira et al., 2011, Eadie et al., 2013, Vaz, 2014, Nègre and Superieur, 2017, Beukelman et al., 2017]. Le pictogramme peut être défini, en CAA, comme un signe graphique schématique dont le signifiant entretient un rapport de ressemblance plus ou moins fort avec le signifié, au contraire des signes linguistiques phoniques ou graphiques dont la forme du stimulus est arbitraire et indépendante de celle du référent : il met en place une représentation iconique de l'information qui est plus facilement interprétable [Norré et al., 2020]. Nous nous concentrons ici sur l'automatisation de la communication via des pictogrammes, comme illustré dans le lien suivant. Les pictogrammes sont couramment utilisés dans les deux sens de la communication :

- la génération vocale à partir de pictogrammes : elle permet de composer un message à partir d'un ensemble de pictogrammes afin de les associer entre eux pour qu'une synthèse vocale énonce le message au destinataire. La génération via logiciel existe déjà sur le marché, de manière simplifiée (un mot correspond à un pictogramme).

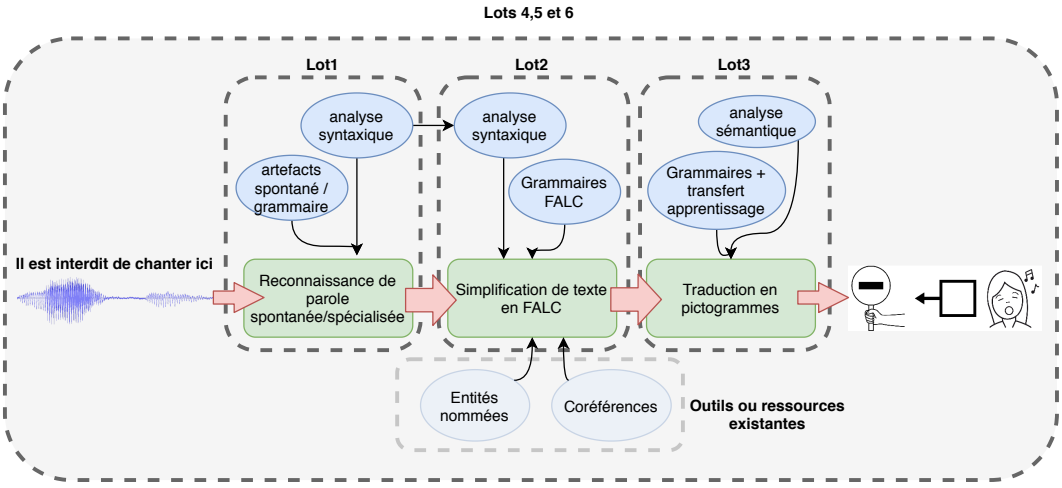


FIGURE 10.1 – Vue globale du projet.

- la génération de pictogrammes à partir du langage naturel oral : associer des pictogrammes au discours correspondant est essentiel au bain langagier.

  Cette association se complique lorsque l'on souhaite projeter une représentation textuelle complexe sous la forme d'un ensemble de pictogrammes. Comme il s'agit de la transposition d'un énoncé linguistique en une représentation visuelle complexe, il ne faut pas seulement associer un pictogramme à un mot, mais plus largement trouver comment des représentations linguistiques peuvent être représentées visuellement [Sevens, 2018].

Dans ce projet, nous nous intéressons spécifiquement à la génération de pictogrammes à partir du langage naturel oral. Notre objectif est d'étudier et de proposer des solutions permettant de réaliser une projection du langage naturel vers un ensemble de pictogrammes, de manière automatique et à partir de la voix. Les principales étapes du projet sont présentées dans la figure 10.1. L'objectif est de parvenir à tout prix à proposer des approches hybrides permettant de lier les avancées en termes d'apprentissage machine aux connaissances linguistiques expertes.

Progrès par rapport à l'état de l'art

Actuellement, les recherches portant sur la TA de la parole vers des pictogrammes n'en sont encore qu'à leurs balbutiements. De plus, les évaluations de

ces approches (à partir de la parole) avec des groupes réels sont pour l'instant inexistantes dans la littérature.

Notons Helpicto de l'entreprise Equadex, une application de génération de pictogrammes à partir de la parole datant de 2017. Aidés par Microsoft, ils ont pu bénéficier de ses API et procéder à l'assemblage de briques logicielles. L'application effectue un mapping entre des mots et des pictogrammes sans prendre en considération les aspects linguistiques complexes abordés au cœur de PROPICTO : par exemple, "la télé marche" est traduite par une télé qui est en train de courir. Au-delà de l'aspect fonctionnel, autant que l'on en sache Helpicto n'a été accompagné d'aucune évaluation scientifique.

Au niveau de la recherche, quelques rares études ont abordé la projection de texte vers des pictogrammes [Mihalcea, 2008, Takasaki and Mori, 2011, Vandeghinste and Schuurman, 2014, Sevens et al., 2015, Vandeghinste et al., 2017, Sevens et al., 2017a, Sevens et al., 2017b]. La plupart d'entre elles sont issues du projet européen *able to include* qui a défriché le sujet en visant à améliorer les conditions de vie des personnes atteintes de troubles du développement intellectuel. Cependant, leur approche s'éloigne des cas d'usage écologiques que nous souhaitons aborder : elles se limitent en effet à une traduction en pictogrammes dans le cadre d'un système de messagerie. Un autre problème est que les pictogrammes proposés dans *able to include* sont, dans la pratique, peu utilisés en France ou en Suisse. Une démonstration de leur système est disponible sur leur site pour le néerlandais, l'anglais et l'espagnol, qui représente l'un des rares systèmes de ce type en ligne à traduire des phrases complètes dans des suites de pictogrammes (voir aussi `https://www.pictotraductor.com/`, pour l'espagnol). Un autre travail intéressant à évoquer dans le cadre de PROPICTO est le récent projet ANR CLEAR [Grabar, 2018], qui est dédié à la simplification des textes médicaux en français.

Dans PROPICTO, notre choix s'orientera plutôt vers des jeux de pictogrammes très utilisés en France ou en Suisse comme Makaton et ARASAAC. Ce choix entraîne des différences méthodologiques par rapport aux travaux de [Vandeghinste et al., 2017], en particulier pour le traitement de la négation et des expressions polylexicales, très importantes dans le domaine médical. Nous souhaitons également élargir le champ d'application des pictogrammes au langage naturel oral, d'abord à des domaines spécialisés (médical), puis au domaine spontané du quotidien et des instituts.

Un autre enjeu dans ce projet est celui de la transmission d'informations en contexte multilingue [Mihalcea, 2008]. En effet, les outils de TA ne couvrent

pas la totalité des langues du monde (presque 7000) et il peut être parfois difficile pour un locuteur de transmettre un message à quelqu'un avec qui il ne partage aucun bagage langagier. Or, certaines situations comme une hospitalisation, des demandes administratives ou la scolarisation d'un enfant nécessitent de faire passer des informations de la manière la plus claire et compréhensible possible. Ces personnes se retrouvent dans une situation de barrière communicationnelle : un outil de TA de la parole vers des pictogrammes introduirait une nouvelle modalité plus iconique pour illustrer le propos et améliorer l'intercompréhension.

Pour atteindre ses objectifs, PROPICTO mutualisera le savoir-faire d'équipes multi-disciplinaires avec des connaissances en accessibilité, linguistique, traductologie et traitement automatique des langues afin d'étudier et de proposer des méthodes de communication en français utilisables quotidiennement et évaluées avec des groupes cibles liés à la médiation, au médical et au handicap, notamment les travaux préliminaires de TIM [Mutal et al., 2019, Norré et al., 2020, Gerlach et al., 2020] et GETALP [Schwab, 2018, Vaschalde et al., 2018a, Vaschalde et al., 2018b] autour du handicap et de la projection de la langue naturelle vers des pictogrammes (voir ci-dessous). Ces projets ont permis d'identifier les nombreux verrous scientifiques : il améliorera ainsi l'état de l'art et les connaissances concernant la RAP spontanée, l'analyse syntaxique de la parole spontanée, la simplification de texte en FALC, la TA en pictogrammes avec de faibles ressources.

Finalement, PROPICTO s'articulera autour de trois études de cas permettant une progression des difficultés : nous commencerons avec des 1) données spécialisées dérivées du projet BabelDr, puis nous nous dirigerons vers des 2) données spontanées issues de corpus existants (ESLO, Orfeo) et nous finirons avec des 3) données écologiques : il s'agit d'un volet important de PROPICTO avec la récolte et l'annotation d'un corpus dédié afin de promouvoir la recherche sur la TA de la parole vers des pictogrammes. Il nécessitera aussi des ressources particulières pour la représentation des entités nommées et des termes spécialisés (noms de maladie, etc.).

10.2 Conclusion et perspectives

10.2.1 Conclusion

Dans ce manuscrit nous avons synthétisé mes contributions durant ces dix dernières années. Mes recherches se sont principalement articulées autour

de deux axes : la reconnaissance automatique de la parole et les mesures de confiance. Les travaux présentés ont été menés en partenariat avec différents membres de l'équipe GETALP et plus particulièrement : Laurent Besacier, François Portet, Didier Schwab, Michel Vacher. Ils ont aussi été portés par plusieurs doctorants que j'ai co-encadrés : Ngoc Quang Luong (thèse soutenue), Ngoc Tien Le (thèse soutenue), Zied Elloumi (thèse soutenue), Loïc Vial (thèse soutenue), Lucia Ormaechea, Solène Evain, Hang Le.

Au niveau de la reconnaissance automatique de la parole nous avons proposé des méthodes permettant de rendre les systèmes plus robustes dans le cadre d'appartements intelligents où les conditions d'enregistrement sont distantes, réalisées avec des micros classiques, en présence de bruit ambiant. Dans ces conditions, nous avons fait la preuve qu'il était envisageable d'intégrer des solutions de reconnaissance de la parole utilisant des ressources locales et destinées à la domotique. Nous avons également décrit nos travaux, relatifs à la traduction automatique de la parole, en montrant comment nous pouvions cascader efficacement un système de reconnaissance de la parole avec un système de traduction automatique. Enfin nous avons évoqué nos contributions autour des systèmes de reconnaissance de la parole guidés par des grammaires et plus récemment des approches basées sur des modèles auto-supervisés.

Si en 2010 le domaine de la reconnaissance de la parole n'était encore qu'un outil réservé aux spécialistes et pratiquement ignoré du grand public, il a connu, depuis, de grandes mutations avec l'avènement des réseaux de neurones profonds et l'accroissement, tant des capacités de calcul que d'acquisition et de stockage des données. La reconnaissance de la parole fait désormais partie de notre quotidien. Les systèmes actuels ont réussi à atteindre des taux d'erreur acceptables pour diverses utilisations industrielles : nous avons d'ailleurs présenté quelques transferts industriels, dont notamment un SRAP pour le CHU de Genève. Cependant, peut-on dire aujourd'hui que la reconnaissance de la parole est maîtrisée ? Pas vraiment : certes les tâches de transcription de *broadcast news* ou de parole lue sur des langues bien dotées ne posent plus vraiment de soucis ... et l'humain peinerait même à concurrencer la qualité des transcriptions. Mais nombre d'obstacles doivent encore être surmontés :

- Sur les langues bien dotées, les accents sont encore un problème non résolu et cela est directement lié à la faible quantité de données.
- Les conditions bruitées dégradent encore fortement les systèmes de reconnaissance : il suffit de regarder les résultats des dernières campagnes Schime où la lecture de transcriptions n'aide sans doute pas beaucoup à comprendre ce qui a été prononcé.

- Du côté des locuteurs multiples avec un seul micro : pour l'humain cette tâche est relativement aisée ... mais pour les systèmes de transcription automatique, c'est problématique.
- Le bilan est plus catastrophique sur les langues peu dotées qui qui sont majoritaires en terme de nombre : les systèmes sont souvent peu performants, voire inexistants.
- Du point de vue des métriques, le WER est toujours majoritairement utilisé, mais bien qu'il soit très objectif au niveau du mot, il n'est pas du tout informatif sur la qualité sémantique du message produit. Selon les mots substitués, l'impact sur la compréhension de la transcription peut être important. Cela est lié au fait que les systèmes intègrent encore peu de contexte large.

Il ne faut pas nier, cependant, les grands progrès réalisés. Les systèmes de RAP adaptés à une tâche particulière sont désormais envisageables dans de nombreux domaines. Par ailleurs, les approches auto-supervisées n'en sont qu'à leurs débuts et laissent envisager des progrès qu'on n'imagine peut être pas encore pleinement, que ce soit sur les langues peu dotées ou dans les conditions difficultueuses.

En ce qui concerne les mesures de confiance, nous avons proposé des mesures efficaces à la fois pour la reconnaissance automatique de la parole et la traduction automatique. Dans les deux cas, elles s'articulent autour de paramètres extraits à différents niveaux du système et sont fusionnées avec un classifieur à base de CRF. Dans un second temps, dédié à de la traduction automatique de la parole, nous avons proposé un cadre formel permettant de joindre les mesures issues des deux systèmes. Nous avons également proposé des mesures permettant de prédire l'indexabilité d'une transcription ou encore les performances d'un SRAP en fonction des données que l'on possède. Ces approches nous enseignent de manière générale que des informations externes aux systèmes sont toujours complémentaires et informatives. Avec les réseaux de neurones profonds, les mesures de confiance ont changé de paradigme : sur les anciens systèmes de TA ou de RAP, des paramètres pouvaient être extraits à différents niveaux (graphique, acoustique, linguistique ...). Dans les systèmes à base de réseaux de neurones profonds, les probabilités de sortie du modèle sont souvent interprétées directement comme des mesures de confiance : toutefois il est apparu que cette approche n'était pas exempte de défauts en raison d'une confiance souvent surestimée [Guo et al., 2017]. Actuellement, les mesures de confiance pour les réseaux de neurones profonds font l'objet de nombreuses

recherches : les approches les plus courantes visent à observer le comportement des couches en fonction des erreurs et d'apprendre à un classifieur à reconnaître des comportements typiques d'erreurs. Les recherches autour des mesures de confiance sont cruciales pour les systèmes de bout-en-bout car ces derniers peuvent dans certaines situations présenter des résultats bien plus erratiques que les systèmes classiques (certaines perturbations du signal d'entrée induisant des sorties extravagantes). Nous avons observé ce type de comportement dans le cadre du projet TRIDAN avec de la traduction neuronale sur des sorties OCR : dans certains cas le système "inventait" des traductions en raison d'erreurs au niveau des caractères. Plus dramatiquement, nous pouvons prendre comme exemple les accidents liés aux systèmes de navigation autonome mis sur le marché par TESLA qui mettent en évidence une carence d'auto-évaluation du système.

10.2.2 Perspectives

Perspectives à court et moyen terme

Actuellement je co-encadre Solène Evain avec Solange Rossato et François Portet. Sa thèse porte sur l'étude de la reconnaissance automatique de la parole spontanée. Suite au départ de Laurent Besacier de l'équipe GETALP, j'ai repris la co-direction avec Didier Schwab du travail de thèse de Hang Le, objet d'un partenariat avec Facebook sur la traduction de la parole multilingue avec des systèmes de bout-en-bout.

Pour les quatre années à venir, deux projets ANR seront le fil conducteur de mes activités de recherche :

- Le projet ANR international PROPICTO, décrit précédemment, s'inscrit dans le domaine des humanités numériques, à la frontière entre TAL et sciences humaines et sociales. PROPICTO abordera plusieurs domaines du TALN et propulsera de nouvelles tâches autour de la projection de l'oral vers des pictogrammes, en proposant des méthodes d'évaluation rigoureuses et la création de corpus de référence. Par ailleurs, l'un des autres enjeu majeur de PROPICTO est la volonté d'injecter des connaissances linguistiques expertes dans des approches *machine learning*. D'un point de vue sociétal, les méthodes découlant de ce projet seront d'une grande aide pour différentes structures et permettront à de nombreuses personnes en situation de handicap, allophones, d'avoir la possibilité de développer leur communication, leur autonomie et leur accès à l'information. PROPICTO

implique une grande partie de l'équipe GETALP (Maximin Coavoux, Emmanuelle Esperança-Rodier, Jerôme Goulian, François Portet, Solange Rossato, Didier Schwab) et nous espérons qu'il sera un projet fédérateur. Ce projet finance 3 thèses qui débuteront en novembre 2021 autour des sujets suivants :

  - La reconnaissance de la parole spontanée, mais contrainte par son environnement (médical, ou aidants auprès de personnes en situation de handicap)
  - L'analyse syntaxique de la parole
  - La traduction de la parole vers des pictogrammes

  Ces recherches seront amenées à interagir entre elles et aborderont la problématique de cascade d'erreurs entre différents systèmes (de bout-en-bout ou autres).
- CREAM : c'est un projet ANR porté par Emmanuel Schang de l'université d'Orléans et qui porte sur la mise à disposition d'outils automatiques de traitement des langues pour les langues Créoles. L'équipe de recherche d'Orléans est essentiellement composée de linguistes qui ont conscience de l'aide que pourraient apporter certains outils de TAL pour leurs recherches (alignement, reconnaissance de la parole etc.). Dans le cadre de ce projet nous allons aborder la conception d'outils pour des langues qui sont plus ou moins bien dotées et qui peuvent se rapprocher de langues très dotées. Nous allons étudier comment transférer des connaissances de langues bien dotées vers différents Créoles et explorer des techniques pour réaliser des outils avec des quantités de données restreintes. Dans le cadre de CREAM je vais co-encadrer une thèse qui débutera en novembre 2021 avec Emmanuel Schang. L'objectif sera de proposer une méthodologie pour porter des outils phares du TALN pour les langues Créoles.

Perspectives à plus long terme

À plus long terme, je souhaiterais renforcer les collaborations pluridisciplinaires, notamment avec les linguistes, phonéticiens, psychologues... En effet, le premier défi dans l'articulation TAL/*deep learning* auquel je souhaite contribuer sera l'introduction des connaissances expertes au sein des approches de bout-en-bout où les quantités d'exemples phénoménales se substituent pour l'instant aux expertises : il en découle des capacités de calcul nécessaires qui sont devenues difficilement accessibles. Par exemple, pour reproduire les modèles auto-supervisés de type *wav2vec* pour le français nous avons dû utiliser le

super-calculateur Jean-Zay durant plusieurs semaines et réaliser le travail en collaboration avec 3 laboratoires. La recherche a été rendue très compliquée par ces aspects calculatoires. Par ailleurs, aujourd'hui seuls les grands acteurs industriels de l'informatique (Google, Apple, Facebook, Amazon, Microsoft ...) ont les capacités de calcul suffisantes pour mener sereinement l'apprentissage de leurs modèles. De plus, les problèmes écologiques liés à ces modèles commencent à être soulevés en raison de leur impact carbone [Parcollet and Ravanelli, 2021]. Un moyen d'y pallier serait de mieux mutualiser ces modèles/apprentissages et mettre en œuvre des méthodes permettant de les adapter à faible coût calculatoire. L'intégration de connaissances expertes dans l'aide à la décision des réseaux de neurones peut aider à rendre les modèles plus simples. En lien avec l'intégration de connaissances expertes, j'ai également l'intention de travailler sur l'explicabilité des informations capturées au sein des SRAP à base de réseaux de neurones profonds. Cela donnera des pistes pour des apprentissages plus efficaces et un transfert plus aisé des modèles d'une langue à l'autre. Parallèlement, les approches auto-supervisées me semblent prometteuses : dans cette optique le prochain défi sera de joindre plusieurs modalités dans un apprentissage auto-supervisé. Les plongements auto-supervisés de différentes modalités vers un espace commun permettent d'envisager la mutualisation de connaissances très diverses en termes d'origine et des passages plus aisés d'une modalité à une autre. D'ailleurs, aujourd'hui le concept de *foundation model* commence à émerger [Bommasani et al., 2021], se positionnant comme une suite logique au *deep learning* : les auteurs prévoient que les modèles issus de différents domaines vont commencer à converger et seront adaptables à de nombreuses tâches. Une autre perspective, plus générale, directement liée à celles évoquées précédemment est de m'orienter vers les recherches liées aux handicaps cognitifs. Cet intérêt a été suscité par Didier Schwab auprès duquel j'ai eu l'opportunité de découvrir l'existence d'un milieu pluridisciplinaire - par essence enrichissant - qui accompagne des recherches d'autant plus gratifiantes qu'à ce jour la problématique des handicaps cognitifs est peu couverte par les grands acteurs industriels.

Bibliographie

[Aman et al., 2013] Aman, F., Auberge, V., and Vacher, M. (2013). How affects can perturbe the automatic speech recognition of domotic interactions. In *Proc. of WASSS 2013, Satellite workshop of Interspeech 2013*, pages 1–5, Grenoble, France.

[Aman et al., 2016a] Aman, F., Aubergé, V., and Vacher, M. (2016a). Robustesse de la RAP à la parole expressive âgée vs. typique : contexte de commandes dans un habitat intelligent. In *JEP 2016*. (Submited paper).

[Aman et al., 2016b] Aman, F., Vacher, M., Portet, F., Duclot, W., and Lecouteux, B. (2016b). CirdoX : an On/Off-line Multisource Speech and Sound Analysis Software. In *Language Resources and Evaluation Conference*, Language Resources and Evaluation Conference, pages 1978–1985, Portoroz, Slovenia. ELRA.

[Ardila et al., 2019] Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F., and Weber, G. (2019). Common voice : A massively-multilingual speech corpus.

[Awni, 2018] Awni, A. (2018). https ://awni.github.io/speech-recognition/.

[Baevski et al., 2020] Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations. *arXiv :2006.11477*.

[Bandeira et al., 2011] Bandeira, F. M., Faria, F. P. D., and Araujo, E. B. D. (2011). Quality assessment of inhospital patients unable to speak who use alternative and extended communication. *Einstein, São Paulo, Brésil, 9(4), 2011, p. 477*.

[Bechet et al., 2015] Bechet, F., Favre, B., and Rouvier, M. (2015). "speech is silver, but silence is golden" : improving speech-to-speech translation performance by slashing users input. In *Proceedings of Interspeech 2015*, Dresden, Germany.

[Belinkov and Glass, 2017] Belinkov, Y. and Glass, J. (2017). Analyzing hidden representations in end-to-end automatic speech recognition systems. In *Advances in Neural Information Processing Systems*.

[Bellegarda, 2000] Bellegarda, J. (2000). Exploiting latent semantic information in statistical language modeling. *Proceedings of the IEEE*, 88(8) :1279–1296.

[Bellegarda, 2013] Bellegarda, J. (2013). Large-scale personal assistant technology deployment : the siri experience. In *Proc. Interspeech 2013*, pages 2029–2033.

[Besacier et al., 2012] Besacier, L., Lecouteux, B., Azouzi, M., and Luong, N. (2012). The LIG English to French Machine Translation System for IWSLT 2012. In *9th International Workshop on Spoken Language Translation (IWSLT)*, pages 102–108, Hong Kong, China.

[Besacier et al., 2014] Besacier, L., Lecouteux, B., Luong, N. Q., Hour, K., and Hadjsalah, M. (2014). Word confidence estimation for speech translation. In *Proceedings of The International Workshop on Spoken Language Translation (IWSLT)*.

[Beukelman et al., 2017] Beukelman, D., Mirenda, P., and Superieur, D. (2017). Communication alternative et augmentée : Aider les enfants et les adultes avec des difficultés de communication. *Ouvrage*.

[Bicici, 2013] Bicici, E. (2013). Referential translation machines for quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Association for Computational Linguistics.

[Blatz et al., 2004] Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N. (2004). Confidence estimation for machine translation. In *Proceedings of COLING 2004*.

[Bommasani et al., 2021] Bommasani, R., Hudson, D., Adeli, E., et al. (2021). On the opportunities and risks of foundation models.

[Bonabeau and Théraulaz, 2000] Bonabeau, E. and Théraulaz, G. (2000). L'intelligence en essaim. *Pour la science*, pages 66–73.

[Bontoux, 2008] Bontoux, B. (2008). *Techniques hybrides de recherche exacte et approchée : application à des problèmes de transport*. PhD thesis, Université d'Avignon et des pays de Vaucluse.

[Bouillon et al., 2017] Bouillon, P., Gerlach, J., Spechbach, H., Tsourakis, N., and Halimi, S. (2017). BabelDr vs Google Translate : a user study at Geneva University Hospitals (HUG). In *Proceedings of the 20th Annual Conference of the European Association for Machine Translation*, Prague.

[Bourdin and Navion, 2013] Bourdin, D. and Navion, A. (2013). Mesure de l'efficacité vocale au sein d'une population de chanteurs de human beatbox : Analyse acoustique, aérodynamique et observation comportamentale. *Mémoire d'Orthophonie, Université Claude Bernard Lyon1 - ISTR - Orthophonie.*

[Bouzidi et al., 2017] Bouzidi, K., Elloumi, Z., Besacier, L., Lecouteux, B., and Faouzi BenZeghiba, M. (2017). Traitement des Mots Hors Vocabulaire pour la Traduction Automatique de Document OCRisés en Arabe. In *TALN 2017*, Orléans, France.

[Chang et al., 2008] Chang, H.-l., Pan, Y.-c., and Lee, L.-s. (2008). Latent semantic retrieval of spoken documents over position specific posterior lattices. In *SLT Workshop, 2008. SLT 2008. IEEE*, pages 285–288.

[Chelba et al., 2007] Chelba, C., J. Silva, and Acero, A. (2007). Soft indexing of speech content for search in spoken documents. *Computer Speech and Language*, 21 :458–478.

[Clouet and de Torcy, 2010] Clouet, A. and de Torcy, T. (2010). Le human beatbox : études qualitatives acoustique en video-fibronasoscopie. *Mémoire d'Orthophonie,Université Paris 6.*

[Cox and Dasmahapatra, 2002] Cox, S. and Dasmahapatra, S. (2002). High-level approaches to confidence estimation in speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 10(7) :460–471.

[Dai et al., 2017] Dai, W., Dai, C., Qu, S., Li, J., and Das, S. (2017). Very deep convolutional neural networks for raw waveforms. *International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

[Dalvi et al., 2018] Dalvi, F., Durrani, N., Sajjad, H., and Vogel, S. (2018). Incremental decoding and training methods for simultaneous translation in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, pages 493–499, New Orleans, Louisiana. Association for Computational Linguistics.

[Davis et al., 1952] Davis, K., Biddulph, R., and Balashek, S. (1952). Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, 24(6) :637–642.

[de Torcy et al., 2014] de Torcy, T., Clouet, A., Pillot-Loiseau, C., Vaissiere, J., Brasnu, D., and Crevier-Buchman, L. (2014). A video-fiberscopy study of laryngo-pharyngeal behaviour in the human beatbox. *Logopedics Phoniatrics Vocology, 39, 1, 38-48.*

[Dehak et al., 2010] Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., and Ouellet, P. (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*.

[Devlin et al., 2019] Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT : pre-training of deep bidirectional transformers for language understanding. *to appear in NAACL*.

[Dixon et al., 2011] Dixon, P., Finch, A., Hori, C., and Kashioka, H. (2011). Investigation on the effects of ASR tuning on speech translation performance. In *The proceedings of the International Workshop on Spoken Language Translation (IWSLT 2011)*, San Francisco.

[Duboisdindien, 2014] Duboisdindien, G. (2014). L'interprétation des pictogrammes. statut linguistique et limites de l'utilisation des pictogrammes dans la réhabilitation langagière. - Étude de deux groupes d'enfants âgés de 5 à 6 ans – entraînés versus non entraînés. *Mémoire de recherche de Master de Linguistique Générale et Appliquée*.

[Eadie et al., 2013] Eadie, K., Carlyon, M. J., Stephens, J., and Wilson, M. D. (2013). Communicating in the pre-hospital emergency environment. *Australian Health Review, 37(2), 2013, p. 140-146.*

[Elbayad et al., 2020a] Elbayad, M., Besacier, L., and Verbeek, J. (2020a). Efficient Wait-k Models for Simultaneous Machine Translation. In *Interspeech 2020 - Conference of the International Speech Communication Association*, pages 1461–1465, Shangai (Virtual Conf), China.

[Elbayad et al., 2020b] Elbayad, M., Nguyen, H., Bougares, F., Tomashenko, N., Caubrière, A., Lecouteux, B., Estève, Y., and Besacier, L. (2020b). ON-TRAC Consortium for End-to-End and Simultaneous Speech Translation Challenge Tasks at IWSLT 2020. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 35–43, Seattle, WA, United States. Association for Computational Linguistics.

[Elloumi, 2019] Elloumi, Z. (2019). *Prédiction de performances des systèmes de Reconnaissance Automatique de la Parole*. Theses, Université Grenoble Alpes.

[Elloumi et al., 2018a] Elloumi, Z., Besacier, L., Galibert, O., Kahn, J., and Lecouteux, B. (2018a). ASR performance prediction on unseen broadcast programs using convolutional neurol networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada.

[Elloumi et al., 2018b] Elloumi, Z., Besacier, L., Galibert, O., and Lecouteux, B. (2018b). Analyzing Learned Representations of a Deep ASR Performance Prediction Model. In *Blackbox NLP Workshop and EMLP 2018*, Bruxelles, Belgium.

[Elloumi et al., 2018c] Elloumi, Z., Lecouteux, B., Galibert, O., and Besacier, L. (2018c). Prédiction de performance des systèmes de reconnaissance automatique de la parole à l'aide de réseaux de neurones convolutifs. *Revue TAL*.

[Evain et al., 2019] Evain, S., Contesse, A., Pinchaud, A., Schwab, D., Lecouteux, B., and Henrich Bernardoni, N. (2019). Beatbox sounds recognition using a speech-dedicated HMM-GMM based system. In *MAVEBA 2019 - 11th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, Florence, Italy.

[Evain et al., 2020a] Evain, S., Contesse, A., Pinchaud, A., Schwab, D., Lecouteux, B., and Henrich Bernardoni, N. (2020a). Reconnaissance de parole beatboxée à l'aide d'un système HMM-GMM inspiré de la reconnaissance automatique de la parole. In Benzitoun, C., Braud, C., Huber, L., Langlois, D., Ouni, S., Pogodalla, S., and Schneider, S., editors, *JEP-TALN-RECITAL 2020 - 6e conférence conjointe 33e Journées d'Études sur la Parole, 27e Traitement Automatique des Langues Naturelles, 22e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 208–216, Nancy, France. ATALA.

[Evain et al., 2020b] Evain, S., Lecouteux, B., Portet, F., Estève, I., and Fabre, M. (2020b). Towards Automatic Captioning of University Lectures for French Students who are Deaf. In *ACM SIGACCESS Conference on Computers and Accessibility*, Athènes, Greece.

[Evain et al., 2021a] Evain, S., Lecouteux, B., Schwab, D., Contesse, A., Pinchaud, A., and Henrich Bernardoni, N. (2021a). Human Beatbox Sound Recognition using an Automatic Speech Recognition Toolkit. *Biomedical Signal Processing and Control*, 67 :102468.

[Evain et al., 2021b] Evain, S., Nguyen, H., Le, H., Zanon Boito, M., Mdhaffar, S., Alisamir, S., Tong, Z., Tomashenko, N., Dinarelli, M., Parcollet, T., Allauzen, A., Estève, Y., Lecouteux, B., Portet, F., Rossato, S., Ringeval, F., Schwab, D., and Besacier, L. (2021b). LeBenchmark : A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech. In *Interspeech 2021 : Conference of the International Speech Communication Association*, Brno, Czech Republic.

[Falavigna et al., 2002] Falavigna, D., Gretter, R., and Riccardi, G. (2002). Acoustic and word lattice based algorithms for confidence scores. *Interspeech*, pages 1621–1624.

[Fetter et al., 1996] Fetter, P., Dandurand, F., and Regel-Brietzmann, P. (1996). Word graph rescoring using confidence measures. In *Fourth International Conference on Spoken Language ICSLP*, volume 1, pages 10–13 vol.1.

[Fomicheva et al., 2020] Fomicheva, M., Sun, S., Yankovskaya, L., Blain, F., Guzmán, F., Fishel, M., Aletras, N., Chaudhary, V., and Specia, L. (2020). Unsupervised Quality Estimation for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8 :539–555.

[Galliano et al., 2006] Galliano, S., Geoffrois, E., Gravier, G., Bonastre, J.-F., Mostefa, D., and Choukri, K. (2006). Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In *In Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC 2006)*.

[Galliano et al., 2005] Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.-F., and Gravier, G. (2005). The ester phase ii evaluation campaign for the rich transcription of french broadcast news. In *Interspeech*, pages 1149–1152.

[Garofolo et al., 2000] Garofolo, J., Auzanne, C., and Voorhees, E. (2000). The TREC spoken document retrieval track : A success story. In *in TREC 8*, pages 16–19.

[Gerlach et al., 2020] Gerlach, J., P. Bouillon, and H. Spechbach (2020). A bidirectional translation system for medical diagnostic dialogues using pictographs for patient responses. In *5th Spring Congress of the Swiss Society of General Internal Medicine 2020, à paraître*.

[Geurts et al., 2006] Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 36(1) :3–42.

[Grabar, 2018] Grabar, N. (2018). projet clear (communication, literacy, education, accessibility, readability) financé l'anr sous la référence anr-17-ce19-0016-01.

[Grijalba Ormaechea et al., 2021] Grijalba Ormaechea, L., Bouillon, P., Gerlach, J., Lecouteux, B., Schwab, D., and Spechbach, H. (2021). Reconnaissance vocale du discours spontané pour le domaine médical. In *Technologies du Langage Humain (TLH)*, Paris, France.

[Guo et al., 2017] Guo, C., Pleiss, G., Sun, Y., , and Weinberger, K. Q. (2017). On calibration of modern neural networks. *International Conference on Machine Learning*.

[Guo et al., 2004] Guo, G., Huang, C., Jiang, H., and Wang, R.-H. (2004). A comparative study on various confidence measures in large vocabulary speech recognition. In *2004 International Symposium on Chinese Spoken Language Processing*, pages 9–12.

[Gupta et al., 2015] Gupta, R., Orasan, C., and Genabith, J. (2015). Machine translation evaluation using recurrent neural networks. In *Proceedings Workshop on Machine Translation (WMT), Metrics Shared Task*, Lisbonne, Portugal.

[Hacker et al., 2015] Hacker, K., Anies, M., Folb, B., and Zallman, L. (2015). Barriers to health care for undocumented immigrants : a literature review. *Risk Management and Healthcare Policy*, pages 175–183.

[Hadj Salah et al., 2018] Hadj Salah, M., Vial, L., Blanchon, H., Zrigui, M., Lecouteux, B., and Schwab, D. (2018). Traduction automatique de corpus en anglais annotés en sens pour la désambiguïsation lexicale d'une langue moins bien dotée, l'exemple de l'arabe. In *25e conférence sur le Traitement Automatique des Langues Naturelles*, Rennes, France.

[Hakkani-Tür et al., 2005] Hakkani-Tür, D., Tur, G., Ricardi, G., and Kim, H. (2005). Error prediction in spoken dialog : from signal-to-noise ratio to semantic confidence scores. In *IEEE International Conference on Acoustics, Speech and Language Processing*, volume I, pages 1041–1044.

[Han et al., 2013] Han, A., Lu, Y., Wong, D., Chao, L., He, L., and Xing, J. (2013). Quality estimation for machine translation using the joint method of evaluation criteria and statistical modeling. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Association for Computational Linguistics.

[Hansen et al., 2005] Hansen, J., Huang, R., Zhou, B., Seadle, M., Deller, J., Gurijala, A., Kurimo, M., and Angkititrakul, P. (2005). Speechfind : Advances in spoken document retrieval for a national gallery of the spoken word. *Speech and Audio Processing, IEEE Transactions on*, 13(5) :712–730.

[Hinton et al., 2012] Hinton, G., Deng, L., Yu, D., Dahl, G., M., A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., , and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition : The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6) :82–97.

[Horndasch et al., 2016] Horndasch, A., Kaufhold, C., and Nöth., E. (2016). How to add word classes to the kaldi speech recognition toolkit. *In International Conference on Text, Speech, and Dialogue*, page 486–494.

[Huang et al., 2014] Huang, X., Baker, J., and Reddy, R. (2014). A historical perspective of speech recognition. *Communications of the ACM*, 57 :94–103.

[Itakura and Saito, 1970] Itakura, F. and Saito, S. (1970). A statistical method for estimation of speech spectral density and formant frequencies. *Electronics and communications in Japan*.

[Jalalvand et al., 2016] Jalalvand, S., Negri, M., Turchi, M., de Souza, J. G., Falavigna, D., and Qwaider, M. (2016). Transcrater : a tool for automatic speech recognition quality estimation. *Proceedings of ACL-2016 System Demonstrations. Berlin, Germany : Association for Computational Linguistics*, pages 43–48.

[Kemp and Schaaf, 1997] Kemp, T. and Schaaf, T. (1997). Estimating confidence using word lattices. In *Proc. Eurospeech '97*, pages 827–830, Rhodes, Greece.

[Kim, 2014] Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv :1408.5882*.

[Kirchhoff and Yang, 2005] Kirchhoff, K. and Yang, M. (2005). Improved language modeling for statistical machine translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*.

[Koehn et al., 2007a] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007a). Moses : Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*.

[Koehn et al., 2007b] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007b). Moses : Open source toolkit for statistical machine translation. *ACL*.

[Kuhn and De Mori, 1990] Kuhn, R. and De Mori, R. (1990). A cache-based natural language model for speech recognition. *Journal : IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12 :570–583.

[Kumar and Sarawagi, 2019] Kumar, A. and Sarawagi, S. (2019). Calibration of encoder decoder models for neural machine translation.

[Kurimo and Turunen, 2005] Kurimo, M. and Turunen, V. (2005). Retrieving speech correctly despite the recognition errors. In *2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*.

[Lafferty et al., 2001] Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields : Probabilistic models for segmenting et labeling sequence data. In *Proceedings of ICML-01*.

[Lavergne et al., 2010] Lavergne, T., Cappé, O., and Yvon, F. (2010). Practical very large scale crfs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

[Le et al., 2021] Le, H., Barbier, F., Nguyen, H., Tomashenko, N., Mdhaffar, S., Gahbiche, S., Bougares, F., Lecouteux, B., Schwab, D., and Estève, Y. (2021). ON-TRAC' systems for the IWSLT 2021 low-resource speech translation and multilingual speech translation shared tasks. In *International Conference on Spoken Language Translation (IWSLT)*, Bangkok (virtual), Thailand.

[Le et al., 2020] Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2020). Flaubert : Unsupervised language model pre-training for french.

[Le, 2018] Le, N. (2018). *Advanced Quality Measures for Speech Translation*. Theses, Université Grenoble Alpes.

[Le et al., 2016a] Le, N., Lecouteux, B., and Besacier, L. (2016a). Joint ASR and MT Features for Quality Estimation in Spoken Language Translation. In *International Workshop on Spoken Language Translation*, Seattle, United States.

[Le et al., 2017] Le, N., Lecouteux, B., and Besacier, L. (2017). Disentangling ASR and MT Errors in Speech Translation. In *MT Summit 2017*, Nagoya, Japan.

[Le et al., 2018] Le, N., Lecouteux, B., and Besacier, L. (2018). Automatic quality estimation for speech translation using joint ASR and MT features. *Machine Translation*.

[Le et al., 2016b] Le, N., Servan, C., Lecouteux, B., and Besacier, L. (2016b). Better Evaluation of ASR in Speech Translation Context Using Word Embeddings. In *Interspeech 2016*, San-Francisco, United States.

[Lecouteux et al., 2011a] Lecouteux, B., Besacier, L., and Blanchon, H. (2011a). LIG English-French Spoken Language Translation System for IWSLT 2011. In *International Workshop on Spoken Language Translation IWSLT 2011)*, San Francisco, United States.

[Lecouteux et al., 2006a] Lecouteux, B., Linarès, G., Bonastre, J., and Nocera, P. (2006a). Imperfect transcript driven speech recognition. In *Interspeech'06-ICSLP*, Pittburgh, Pensylvania, USA.

[Lecouteux et al., 2008] Lecouteux, B., Linarès, G., Estève, Y., and Gravier, G. (2008). Generalized driven decoding for speech recognition system combination. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Las Vegas, USA.

[Lecouteux et al., 2013] Lecouteux, B., Linares, G., Estève, Y., and Gravier, G. (2013). Dynamic Combination of Automatic Speech Recognition Systems by Driven Decoding. *IEEE Transactions on Audio, Speech and Language Processing*, page na.

[Lecouteux et al., 2007] Lecouteux, B., Linarès, G., Estève, Y., and Mauclair, J. (2007). System Combination by Driven Decoding. In *IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07* , Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, Honolulu, United States.

[Lecouteux et al., 2009a] Lecouteux, B., Linarès, G., and Favre, B. (2009a). Combined low level and high level features for Out-Of-Vocabulary Word detection. In *Interspeech 2009*, Brighton, United Kingdom.

[Lecouteux et al., 2009b] Lecouteux, B., Linares, G., and Favre, B. (2009b). Détection de mots hors-vocabulaire par combinaison de mesures de confiance de haut et bas niveaux. In *MajecSTIC 2009* , Avignon, France.

[Lecouteux et al., 2006b] Lecouteux, B., Linarès, G., Nocera, P., and Bonastre, J.-F. (2006b). Imperfect transcript driven speech recognition. In *Interspeech*, Pittsburgh, United States.

[Lecouteux et al., 2010] Lecouteux, B., Nocera, P., and Linarès, G. (2010). Semantic cache model driven speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, Dallas, United States.

[Lecouteux and Schwab, 2014] Lecouteux, B. and Schwab, D. (2014). Décodage de graphe à l'aide de colonies de fourmis. In *30èmes Journées d'étude de la parole*, page 6, Le mans, France.

[Lecouteux and Schwab, 2015] Lecouteux, B. and Schwab, D. (2015). Ant colony algorithm applied to automatic speech recognition graph decoding. In *Interspeech 2015*, Dresden, Germany.

[Lecouteux et al., 2011b] Lecouteux, B., Vacher, M., and Portet, F. (2011b). Distant Speech Recognition for Home Automation : Preliminary Experimental Results in a Smart Home. In *IEEE SPED 2011*, pages 41–50, Brasow, Romania.

[Lecouteux et al., 2011c] Lecouteux, B., Vacher, M., and Portet, F. (2011c). Distant Speech Recognition in a Smart Home : Comparison of Several Multisource ASRs in Realistic Conditions. In Association, I. S. C., editor, *Interspeech 2011 Florence*, pages 2273–2276, Florence, Italy. 4 pages.

[Lecouteux et al., 2012] Lecouteux, B., Vacher, M., and Portet, F. (2012). Reconnaissance automatique de la parole distante dans un habitat intelligent : méthodes multi-sources en conditions réalistes (Distant Speech Recognition in a Smart Home : Comparison of Several Multisource ASRs in Realistic Conditions) [in French]. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1 : JEP*, pages 657–664, Grenoble, France. ATALA/AFCP.

[Lecouteux et al., 2018] Lecouteux, B., Vacher, M., and Portet, F. (2018). Distant Speech Processing for Smart Home Comparison of ASR approaches in distributed microphone network for voice command. *International Journal of Speech Technology*, 21 :601–618.

[Li et al., 2021] Li, Q., Zhang, Y., Li, B., Cao, L., and Woodland, P. (2021). Residual energy-based models for end-to-end speech recognition.

[Likhomanenko et al., 2021] Likhomanenko, T., Xu, Q., Pratap, V., Tomasello, P., Kahn, J., Avidov, G., Collobert, R., and Synnaeve, G. (2021). Rethinking evaluation in asr : Are our models robust enough ?

[Lowerre, 1976] Lowerre, B. (1976). *The Harpy speech recognition system.* PhD thesis, Carnegie Mellon University.

[Luong, 2014] Luong, N. (2014). *Word Confidence Estimation and Its Applications in Statistical Machine Translation.* Theses, Université de Grenoble.

[Luong et al., 2013a] Luong, N., Besacier, L., and Lecouteux, B. (2013a). Word Confidence Estimation and its Integration in Sentence Quality Estimation for Machine Translation. In *Proceedings of the fifth international conference on knowledge and systems engineering (KSE)*, pages x–x, Hanoi, Vietnam.

[Luong et al., 2014a] Luong, N., Besacier, L., and Lecouteux, B. (2014a). An Efficient Two-Pass Decoder for SMT Using Word Confidence Estimation. In *European Association for Machine Translation (EAMT)*, Dubrovnik, Croatia.

[Luong et al., 2014b] Luong, N., Besacier, L., and Lecouteux, B. (2014b). LIG System for Word Level QE task at WMT14. In *Workshop on Machine Translation (WMT)*, Baltimore, United States.

[Luong et al., 2014c] Luong, N., Besacier, L., and Lecouteux, B. (2014c). Word Confidence Estimation for SMT N-best List Re-ranking. In *Proceedings of the Workshop on Humans and Computer-assisted Translation (HaCaT) during EACL*, Gothenburg, Sweden.

[Luong et al., 2015] Luong, N., Besacier, L., and Lecouteux, B. (2015). Towards Accurate Predictors of Word Quality for Machine Translation : Lessons Learned on French - English and English - Spanish Systems. *Data and Knowledge Engineering*, page 11.

[Luong et al., 2017] Luong, N., Besacier, L., and Lecouteux, B. (2017). Find The Errors, Get The Better : Enhancing Machine Translation via Word Confidence Estimation. *Natural Language Engineering*, 1 :1 – 24.

[Luong et al., 2013b] Luong, N., Lecouteux, B., and Besacier, L. (2013b). LIG System for WMT13 QE Task : Investigating the Usefulness of Features in Word Confidence Estimation for MT. In *8th Workshop on Statistical Machine Translation*, pages 386–391, Sofia, Bulgaria.

[Maaten and Hinton, 2008] Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov) :2579–2605.

[Mauclair, 2006] Mauclair, J. (2006). *Mesures de confiance en traitement automatique de la parole et applications*. PhD thesis, LIUM.

[Mauclair et al., 2006] Mauclair, J., Estève, Y., Petit-Renaud, S., and Deléglise, P. (2006). Automatic detection of well recognized words in automatic speech transcription. In *LREC 2006*, Genoa, Italy.

[Meinshausen and Bühlmann, 2010] Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society : Series B (Statistical Methodology), vol. 72, no 4, p. 417-473.*

[Mihalcea, 2008] Mihalcea, R. et Leong, C. (2008). Toward communicating simple sentences using pictorial representations. *Machine Translation, vol. 22, no. 3.*

[Moreno et al., 2001] Moreno, P., Logan, B., and Raj, B. (2001). A boosting approach for confidence scoring. In *Interspeech, Aalborg, Denmark*, pages 2109–2112.

[Mutal et al., 2019] Mutal, J., Bouillon, P., Gerlach, J., Estrella, P., and Spechbach, H. (2019). Monolingual backtranslation in a medical speech translation system for diagnostic interviews-a nmt approach. In *Proceedings of Machine Translation Summit XVII Volume 2 : Translator, Project and User Tracks*, pages 196–203.

[Nanjo et al., 2006] Nanjo, H., Akita, Y., and Kawahara, T. (2006). Computer assisted speech transcription system for efficient speech archive. In *Western Pacific Acoustic conference*, Seoul, Korea.

[Nguyen et al., 2011] Nguyen, B., Huang, F., and Al-Onaizan, Y. (2011). Goodness : A method for measuring machine translation confidence. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*.

[Nocéra et al., 2002] Nocéra, P., Linarès, G., and Massonié, D. (2002). Phoneme lattice based a* search algorithm for speech recognition. *Text, Speech and Dialogue : 5th International Conference, TSD 2002, Brno, Czech Republic*.

[Norré et al., 2020] Norré, M., Bouillon, P., Gerlach, J., and Spechbach, H. (2020). Evaluating the comprehension of arasaac and sclera pictographs for the babeldr patient response interface. *Barrier Free Conference, BFC*.

[Nègre and Superieur, 2017] Nègre, E. C. and Superieur, D. (2017). Communiquer autrement : Accompagner les personnes avec des troubles de la parole ou du langage. *Ouvrage*.

[Oard et al., 2004] Oard, D. W., Soergel, D., Doermann, D., Huang, X., Murray, G. C., Wang, J., Ramabhadran, B., Franz, M., Gustman, S., Mayfield, J., Kharevych, L., and Strassel, S. (2004). Building an information retrieval test collection for spontaneous conversational speech. In *SIGIR '04*, pages 41–48, New York, USA. ACM.

[Och, 2003] Och, F. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.

[Ogata and Goto, 2005] Ogata, J. and Goto, M. (2005). Speech repair : Quick error correction just by using selection operation for speech input interfaces. In *International Conference on Speech Communication and Technology, Interspeech*, pages 133–136, Lisboa, Portugal.

[Ogawa et al., 2021] Ogawa, A., Tawara, N., Kano, T., and Delcroix, M. (2021). Blstm-based confidence estimation for end-to-end speech recognition. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6383–6387.

[Parcollet and Ravanelli, 2021] Parcollet, T. and Ravanelli, M. (2021). The Energy and Carbon Footprint of Training End-to-End Speech Recognizers. working paper or preprint.

[Paroni et al., 2019] Paroni, A., Henrich Bernardoni, N., Loevenbruck, H., Savariaux, C., Calabrèse, P., Fabre, C., Atallah, I., Baraduc, P., and Gerber, S. (2019). Le human beatbox : d'un langage musical à un outil de rééducation orthophonique. *Actes des Journées de Phonétique Clinique, Mons*.

[Petrov and Klein, 2007] Petrov, S. and Klein, D. (2007). Improved inference for unlexicalized parsing. In *HLT-NAACL*.

[Picart et al., 2015] Picart, B., Brognaux, S., and Dupont, S. (2015). Analysis and automatic recognition of Human BeatBox sounds : A comparative study. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4255–4259, Brisbane, QLD, Australia.

[Potet et al., 2010] Potet, M., Besacier, L., and Blanchon, H. (2010). The lig machine translation system for wmt 2010. In Workshop, A., editor, *Proceedings of the joint fifth Workshop on Statistical Machine Translation and Metrics MATR (WMT2010)*.

[Potet et al., 2012] Potet, M., Esperança-Rodier, E., Besacier, L., and Blanchon, H. (2012). Collection of a large database of french-english smt output corrections. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC)*.

[Potet et al., 2011] Potet, M., Rubino, R., Lecouteux, B., Huet, S., Blanchon, H., Besacier, L., and Lefevre, F. (2011). The LIGA machine translation system for WMT 2011. In *EMNLP 2011 Workshop on statistical machine translation*, Edinburgh , United Kingdom.

[Povey et al., 2011a] Povey, D., Burget, L., Agarwal, M., Akyazi, P., Kai, F., Ghoshal, A., Glembek, O., Goel, N., Karafiát, M., Rastrow, A., Rose, R., Schwarz, P., and Thomas, S. (2011a). The subspace Gaussian mixture model – A structured model for speech recognition . *Computer Speech & Language*, 25(2) :404 – 439.

[Povey et al., 2018] Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., and Khudanpur, S. (2018). Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Proc. of Interspeech*.

[Povey et al., 2011b] Povey, D., Ghoshal, A., Boulianne, G., Goel, N., Hannemann, M., Qian, Y., Schwarz, P., and Stemmer, G. (2011b). The kaldi speech recognition toolkit. In *In IEEE 2011 workshop*.

[Qiu et al., 2021] Qiu, D., He, Y., Li, Q., Zhang, Y., Cao, L., and McGraw, I. (2021). Multi-task learning for end-to-end asr word and utterance confidence with deletion prediction.

[Rahim et al., 1997] Rahim, M., Lee, C., and Juang, B.-H. (1997). Discriminative utterance verification for connected digits recognition. *IEEE J SAP*, 5(3) :266–277.

[Rayner et al., 2016] Rayner, M., Armando, A., Bouillon, P., Ebling, S., Gerlach, J., Halimi, S., Strasly, I., and Tsourakis, N. (2016). Helping domain experts build phrasal speech translation systems. In Quesada, J. F., Martín Mateos, F.-J., and Lopez-Soto, T., editors, *Future and Emergent Trends in Language Technology*, volume 9577, pages 41–52. Springer International Publishing, Cham. Series Title : Lecture Notes in Computer Science.

[Reddy, 1976] Reddy, R. (1976). Speech recognition by machine : A review. *Proceedings of the IEEE 64, 4*.

[Rose et al., 1995] Rose, R., Juang, B., and Lee, C. (1995). A training procedure for verifying string hypotheses in continuous speech recognition. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 281–284 vol.1.

[Salton, 1989] Salton, G. (1989). *Automatic text processing : the transformation, analysis, and retrieval of information by computer*. Addison-Wesley.

[Saraclar, 2004] Saraclar, M. (2004). Lattice-based search for spoken utterance retrieval. In *In Proceedings of HLT-NAACL 2004*, pages 129–136.

[Schneider et al., 2019] Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec : Unsupervised Pre-Training for Speech Recognition. In *Proc. Interspeech 2019*, pages 3465–3469.

[Schönherr et al., 2020] Schönherr, L., Golla, M., Eisenhofer, T., J.W., Kolossa, D., and Holz, T. (2020). Unacceptable, where is my privacy ? exploring accidental triggers of smart speakers. *CoRR*, abs/2008.00508.

[Schwab, 2018] Schwab, D. (2018). The gazeplay project : Open and free eye-trackers games and a community for people with multiple disabilities. *16th International Conference on Computers Helping People with Special Needs,*.

[Schwab et al., 2013] Schwab, D., Goulian, J., and Tchechmedjiev, A. (2013). Désambiguïsation lexicale de textes : éfficacité qualitative et temporelle d ?un algorithme à colonies de fourmis. *journal Traitement Automatique des Langues, vol. 54-1*.

[Schwab et al., 2020] Schwab, D., Trial, P., Vaschalde, C., Vial, L., Esperança-Rodier, E., and Lecouteux, B. (2020). Providing semantic knowledge to a set of pictograms for people with disabilities : a set of links between WordNet and Arasaac : Arasaac-WN. In *LREC*, Marseille, France.

[Schwab et al., 2019] Schwab, D., Trial, P., Vaschalde, C., Vial, L., and Lecouteux, B. (2019). Apporter des connaissances sémantiques à un jeu de pictogrammes destiné à des personnes en situation de handicap : Un ensemble de liens entre Princeton WordNet et Arasaac, Arasaac-WN. In Morin, E., Rosset, S., and Zweigenbaum, P., editors, *26e Conférence sur le Traitement Automatique des Langues Naturelles*, pages 619–622, Toulouse, France. ATALA.

[Sehili et al., 2012] Sehili, M. E. A., Lecouteux, B., Vacher, M., Portet, F., Istrate, D., Dorizzi, B., and Boudy, J. (2012). Sound environment analysis in smart home. In *AmI '12 : International Joint Conference on Ambient Intelligence*, volume 7683, pages 208–223, Pisa, Italy. Springer.

[Senay et al., 2010a] Senay, G., Linarès, G., Lecouteux, B., Oger, S., and M., T. (2010a). Décodage interactif de la parole. In *XXVIIIèmes Journées d'Etude sur la Parole (JEP'2010)*, Mons, Belgium.

[Senay et al., 2010b] Senay, G., Linarès, G., Lecouteux, B., Oger, S., and M., T. (2010b). Transcriber driving strategies for transcription aid system. In *LREC*, Valletta, Malta.

[Serpollet et al., 2007] Serpollet, N., Bergounioux, G., Chesneau, A., and Walter, R. (2007). A large reference corpus for spoken french : Eslo 1 and 2 and its variations. *Université d'Orléans*.

[Servan et al., 2015] Servan, C., Le, N., Luong, N. Q., Lecouteux, B., and Besacier, L. (2015). An Open Source Toolkit for Word-level Confidence Estimation in Machine Translation. In *The 12th International Workshop on Spoken Language Translation (IWSLT'15)*, Da Nang, Vietnam.

[Sevens, 2018] Sevens, L. (2018). Words divide, pictographs unite : Pictograph communication technologies for people with an intellectual disability.

[Sevens et al., 2017a] Sevens, L., Jacobs, G., Vandeghinste, V., Schuurman, I., and Eynde, F. V. (2017a). Improving text-to-pictograph translation through word sense disambiguation. *Conference on Lexical and Computational Semantics*.

[Sevens et al., 2015] Sevens, L., Vandeghinste, V., and Schuurman, I. (2015). Extending a dutch text-to-pictograph to english and spanish. *SLPAT*.

[Sevens et al., 2017b] Sevens, L., Vandeghinste, V., Schuurman, I., and Eynde, F. V. (2017b). Simplified text-to-pictograph translation for people with intellectual disabilities. *NLDB*.

[Siegler, 1999] Siegler, M. A. (1999). *Integration of Continuous Speech Recognition and Information Retrieval for Mutually Optimal Performance*. PhD thesis, Carnegie Mellon University.

[Sinyor et al., 2005] Sinyor, E., McKay, C., Fiebrink, R., McEnnis, D., and Fujinaga, I. (2005). Beatbox classification using ACE. *International Conference on Music Information Retrieval*, page 4.

[Snover et al., 2008] Snover, M., Madnani, N., Dorr, B., and Schwartz, R. (2008). Terp system description. In *MetricsMATR workshop at AMTA*.

[Sokolov et al., 2012] Sokolov, A., Wisniewski, G., and Yvon, F. (2012). Non-linear n-best list reranking with few features. In *Proceedings of AMTA*.

[Sukkar et al., 1996] Sukkar, R., Setlur, A., Rahim, M., and Lee, C.-H. (1996). Utterrance verification of keywords strings using word-based minimum verification error(wb-mve) training. In *ICASSP*.

[Takasaki and Mori, 2011] Takasaki, T. and Mori, Y. (2011). Design and development of a pictogram communication system for children around the world. *Proceedings of the 1st International Conference on Intercultural Collaboration*.

[Tiwari, 2010] Tiwari, V. (2010). MFCC and its applications in speaker recognition. *International Journal on Emerging Technologies*, pages 19–22.

[Ueffing et al., 2003] Ueffing, N., Macherey, K., and Ney, H. (2003). Confidence measures for statistical machine translation. In *Proceedings of the MT Summit IX*.

[Ueffing and Ney, 2007] Ueffing, N. and Ney, H. (2007). Word-level confidence estimation for machine translation. In *Computational Linguistics*.

[Vacher et al., 2019] Vacher, M., Aman, F., Rossato, S., Portet, F., and 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* , 12(2) :8 :1–8 :25.

[Vacher et al., 2016a] Vacher, M., Bouakaz, S., Bobillier-Chaumon, M., Aman, F., Khan, R., Bekkadja, S., Portet, F., Guillou, E., Rossato, S., and Lecouteux, B. (2016a). The CIRDO Corpus : Comprehensive Audio/Video Database of Domestic Falls of Elderly People. In *10th International Conference on Language Resources and Evaluation (LREC 2016)*, 10th International Conference on Language Resources and Evaluation (LREC 2016), pages 1389–1396, Portoroz, Slovenia. ELRA, ELRA.

[Vacher et al., 2015a] Vacher, M., Caffiau, S., Portet, F., Meillon, B., Roux, C., Elias, E., Lecouteux, B., and Chahuara, P. (2015a). Evaluation of a

context-aware voice interface for Ambient Assisted Living : qualitative user study vs. quantitative system evaluation. *ACM Transactions on Accessible Computing* , 7(issue 2) :5 :1–5 :36.

[Vacher et al., 2013a] Vacher, M., Chahuara, P., Lecouteux, B., Istrate, D., Portet, F., Joubert, T., Sehili, M. E. A., Meillon, B., Bonnefond, N., Fabre, S., Roux, C., and Caffiau, S. (2013a). The SWEET-HOME Project : Audio Technology in Smart Homes to improve Well-being and Reliance. In *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'13)*, pages 7298–7301, Osaka, Japan.

[Vacher et al., 2011] Vacher, M., Istrate, D., Portet, F., Joubert, T., Chevalier, T., Smidtas, S., Meillon, B., Lecouteux, B., Sehili, M., Chahuara, P., and Méniard, S. (2011). The SWEET-HOME Project : Audio Technology in Smart Homes to improve Well-being and Reliance. In in Medicine, I. E. and Society, B., editors, *33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2011)*, pages 5291–5294, Boston, United States. 4 pages.

[Vacher et al., 2016b] Vacher, M., Lecouteux, B., Aman, F., Portet, F., and Rossato, S. (2016b). Acquisition et reconnaissance automatique d'expressions et d'appels vocaux dans un habitat. In *JEP-TALN-RECITAL 2016*, pages 28–36, Paris, France.

[Vacher et al., 2015b] Vacher, M., Lecouteux, B., Aman, F., Rossato, S., and Portet, F. (2015b). Recognition of Distress Calls in Distant Speech Setting : a Preliminary Experiment in a Smart Home. In *6th Workshop on Speech and Language Processing for Assistive Technologies*, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 1–7, Dresden, Germany. SIG-SLPAT.

[Vacher et al., 2014a] Vacher, M., Lecouteux, B., Chahuara, P., Portet, F., Meillon, B., and Bonnefond, N. (2014a). The Sweet-Home speech and multimodal corpus for home automation interaction. In *The 9th edition of the Language Resources and Evaluation Conference (LREC)*, pages 4499–4506, Reykjavik, Iceland.

[Vacher et al., 2013b] Vacher, M., Lecouteux, B., Istrate, D., Joubert, T., Portet, F., Sehili, M., and Chahuara, P. (2013b). Evaluation of a Real-Time Voice Order Recognition System from Multiple Audio Channels in a Home. In *the 14rd Annual Conference of the International Speech Communication Association, Interspeech*, pages 2062–2064, Lyon, France.

[Vacher et al., 2013c] Vacher, M., Lecouteux, B., Istrate, D., Joubert, T., Portet, F., Sehili, M., and Chahuara, P. (2013c). Experimental Evaluation of

Speech Recognition Technologies for Voice-based Home Automation Control in a Smart Home. In *4th Workshop on Speech and Language Processing for Assistive Technologies*, pages 99–105, Grenoble, France.

[Vacher et al., 2012] Vacher, M., Lecouteux, B., and Portet, F. (2012). Recognition of Voice Commands by Multisource ASR and Noise Cancellation in a Smart Home Environment. In *EUSIPCO (European Signal Processing Conference)*, pages 1663–1667, Bucarest, Romania.

[Vacher et al., 2014b] Vacher, M., Lecouteux, B., and Portet, F. (2014b). Multichannel Automatic Recognition of Voice Command in a Multi-Room Smart Home : an Experiment involving Seniors and Users with Visual Impairment. In *Interspeech 2014*, pages 1008–1012, Singapour, Singapore.

[Vacher et al., 2015c] Vacher, M., Lecouteux, B., and Portet, F. (2015c). On Distant Speech Recognition for Home Automation. In Andreas HOLZINGER, M. Z. and ROECKER, C., editors, *Lecture Notes in Computer Science*, volume 8700 of *Smart Health : Open Problems and Future Challenges*, pages 161–188. Springer. The official version of this draft is available at Springer via http ://dx.doi.org/10.1007/978-3-319-16226-3_7.

[Vacher et al., 2015d] Vacher, M., Lecouteux, B., Serrano-Romero, J., Ajili, M., Portet, F., and Rossato, S. (2015d). Speech and Speaker Recognition for Home Automation : Preliminary Results. In *8th International Conference Speech Technology and Human-Computer Dialogue "SpeD 2015"*, Proceedings of the 8th International Conference Speech Technology and Human-Computer Dialogue, pages 181–190, Bucarest, Romania. IEEE.

[Vacher et al., 2014c] Vacher, M., Portet, F., Aman, F., Lecouteux, B., Rossato, S., and Auberge, V. (2014c). Reconnaissance automatique de la parole dans les habitats intelligents : application à l'assistance à domicile. In des Technologies pour l'Autonomie et de Gérontechnologie (SFTAG), S. F., editor, *4e Journées Annuelles de la Société Française des Technologies pour l'Autonomie et de Gérontechnologie*, Actes JASFTAG 2014, pages 38–41, Paris, France.

[Vacher et al., 2013d] Vacher, M., Portet, F., Lecouteux, B., and Golanski, C. (2013d). Speech analysis for Ambient Assisted Living : technical and user design of a vocal order system. In Hu, F., editor, *Telhealthcare Computing and Engineering : Principles and Design*, pages 607–638. CRC Press, Taylor and Francis Group.

[van der Werff, 2012] van der Werff, L. (2012). *Evaluation of noisy transcripts for spoken document retrieval*. PhD thesis, University of Twente, Enschede, Netherlands.

[Vandeghinste and Schuurman, 2014] Vandeghinste, V. and Schuurman, I. (2014). Linking pictographs to synsets : Sclera2cornetto. *LREC*.

[Vandeghinste et al., 2017] Vandeghinste, V., Schuurman, I., Sevens, L., and Eynd, F. V. (2017). Translating text into pictographs. *Natural Language Engineering*.

[Vaschalde et al., 2018a] Vaschalde, C., Lecouteux, B., and Schwab, D. (2018a). Génération de pictogrammes à partir de la parole spontanée pour la mise en place d'une communication médiée. In *50 ans de linguistique sur corpus oraux : Apports à l'étude de la variation*, Orléans, France.

[Vaschalde et al., 2018b] Vaschalde, C., Trial, P., Esperança-Rodier, E., Schwab, D., and Lecouteux, B. (2018b). Automatic pictogram generation from speech to help the implementation of a mediated communication. In *Conference on Barrier-free Communication*, Geneva, Switzerland.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *NIPS*.

[Vaz, 2014] Vaz, I. (2014). Visual symbols in healthcare settings for children with learning disabilities and autism spectrum disorder. *British Journal of Nursing 22(3), 2013, p. 156-159*.

[Vela and Tan, 2015] Vela, M. and Tan, L. (2015). Predicting machine translation adequacy with document embeddings. In *Proceedings Workshop on Machine Translation (WMT), Metrics Shared Task*, Lisbonne, Portugal.

[Vial, 2020] Vial, L. (2020). *Modèles neuronaux joints de désambiguïsation lexicale et de traduction automatique*. Theses, Université Grenoble Alpes [2020-....].

[Vial et al., 2017a] Vial, L., Lecouteux, B., and Schwab, D. (2017a). Représentation vectorielle de sens pour la désambiguïsation lexicale à base de connaissances. In *24ème Conférence sur le Traitement Automatique des Langues Naturelles*, Orléans, France.

[Vial et al., 2017b] Vial, L., Lecouteux, B., and Schwab, D. (2017b). Sense Embeddings in Knowledge-Based Word Sense Disambiguation. In *12th International Conference on Computational Semantics*, Montpellier, France.

[Vial et al., 2017c] Vial, L., Lecouteux, B., and Schwab, D. (2017c). Uniformisation de corpus anglais annotés en sens. In *24ème Conférence sur le Traitement Automatique des Langues Naturelles*, Orléans, France.

[Vial et al., 2018a] Vial, L., Lecouteux, B., and Schwab, D. (2018a). Approche supervisée à base de cellules LSTM bidirectionnelles pour la désambiguïsation

lexicale. In *25e conférence sur le Traitement Automatique des Langues Naturelles*, Rennes, France.

[Vial et al., 2018b] Vial, L., Lecouteux, B., and Schwab, D. (2018b). UFSAC : Unification of Sense Annotated Corpora and Tools. In *Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

[Vial et al., 2019a] Vial, L., Lecouteux, B., and Schwab, D. (2019a). Approche supervisée à base de cellules LSTM bidirectionnelles pour la désambiguïsation lexicale. *Revue TAL*.

[Vial et al., 2019b] Vial, L., Lecouteux, B., and Schwab, D. (2019b). Compression de vocabulaire de sens grâce aux relations sémantiques pour la désambiguïsation lexicale. In *TALN 2019 (Conférence sur le Traitement Automatique des Langues Naturelles)*, Toulouse, France.

[Vial et al., 2019c] Vial, L., Lecouteux, B., and Schwab, D. (2019c). Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. In *Global Wordnet Conference*, Wroclaw, Poland.

[Vial et al., 2019d] Vial, L., Lecouteux, B., Schwab, D., Le, H., and Besacier, L. (2019d). The LIG system for the English-Czech Text Translation Task of IWSLT 2019. In *IWSLT (16th International Workshop on Spoken Language Translation)*, Hong-Kong, China.

[Vilar et al., 2006] Vilar, D., Xu, J., D'Haro, L. F., and Ney, H. (2006). Error analysis of statistical machine translation output. In *Proceedings of LREC 2006*, Genoa, Italy.

[Vintsiuk, 1968] Vintsiuk, T. (1968). Word recognition from speech with the use of dynamic programming. *Cybernetics, oscow*, 1 :15–22.

[Wessel et al., 1998] Wessel, F., Wessel, F., Macherey, K., and Schluter, R. (1998). Using word probabilities as confidence measures. In Macherey, K., editor, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 225–228.

[Wessel et al., 2000] Wessel, F., Wessel, F., Schluter, R., and Ney, H. (2000). Using posterior word probabilities for improved speech recognition. In Schluter, R., editor, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP '00*, volume 3, pages 1587–1590 vol.3.

[Whitney, 1971] Whitney, A. (1971). A direct method of nonparametric measurement selection. *IEEE Transactions on Computers*, C-20(9) :1100–1103.

[Whittaker et al., 2002] Whittaker, S., Hirschberg, J., Amento, B., Stark, L., Bacchiani, M., Isenhour, P., and Gary, S. (2002). Scanmail : a voicemail interface that makes speech browsable, readable and searchable. In *Proceedings of CHI2002*, pages 275–282. ACM Press.

[Woodward et al., 2020] Woodward, A., Bonnín, C., Masuda, I., Varas, D., Bou-Balust, E., and Riveiro, J. (2020). Confidence measures in encoder-decoder models for speech recognition. In *Interspeech'20*.

[Young, 1994] Young, S. R. (1994). Recognition confidence measures : Detection of misrecognitions and out-of-vocabulary words. *Proc. of International Conference on Acoustics, Speech and Signal Processing*.

Annexe A

Glossaire

- **AFSR** Association Française du syndrome de Rett
- **ARASAAC** Aragonese Center of Augmentative and Alternative Communication
- **ANR** Agence Nationale de la Recherche
- **BER** Boxeme Error Rate
- **BPE** Byte Pair Encoding
- **CAA** Communication Alternative et Augmentée
- **CHU** Centre Hospitalier Universitaire
- **CNN** Convolutionnal Neural Network
- **CNU** Conseil National des Universités
- **CREAM** projet ANR : documentation de langues CREoles Assistée par la Machine
- **DDA** Driven Decoding Algorithm
- **DL** Désambiguïsation Lexicale
- **DNN** Deep Neural Network
- **EEAP** Établissement pour Enfants et Adultes Polyhandicapé
- **EER** Equal Error Rate
- **EM** Espérance/Maximisation
- **CRF** Conditionnal Random Fields
- **DER** Domotic Error Rate
- **FALC** Facile À Lire et à Comprendre
- **FLUE** French Language Understanding Evaluation

- **fMLLR** feature Maximum Likelihood Linear Regression
- **FST** Finite State Transducer
- **GETALP** Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole
- **GMM** Gaussian Mixture Model
- **HMM** Hidden Markov Model
- **HUG** Hôpitaux Universitaires de Genève
- **ICASSP** International Conference on Acoustics, Speech and Signal Processing
- **IME** Institut Médico Éducatif
- **ISCA** International Speech Communication Association
- **JEP** Journées d'Etude de la Parole
- **KALDI** Boîte à outil *open source* destinée à la reconnaissance automatique de la parole
- **LIA** Laboratoire Informatique d'Avignon
- **LIG** Laboratoire Informatique de Grenoble
- **LIUM** Laboratoire Informatique de l'Université du Maine
- **LM** Language Model
- **LPC** Linear Predective Coding
- **LRT** Likelihood ratio testing
- **LSA** Latent Semantic Analysis
- **LSTM** Long Short Term Memory
- **MAE** Mean Absolute Error
- **MHV** Mot Hors Vocabulaire
- **MLLR** Maximum Likelihood Linear Regression
- **NLP** Natural Language Processing
- **OCR** Optical Character Recognition
- **ONU** Organisation des Nations Unies
- **PER** Prediction Error Rate
- **POS** Part Of Speech
- **PROPICTO** PRojection du langage Oral vers des unités PICTOgraphiques

- **RAP** Reconnaissance Automatique de la Parole
- **ROVER** Recognizer Output Voting Error Reduction
- **RT** Real Time
- **SAT** Speaker Adaptative Training
- **SemER** Semantic Error Rate
- **SFS** Sequential Forward Selection
- **s-GMM** Subspace Gaussian Mixture Model
- **SBS** Sequential Backward Selection
- **SHS** Sciences Humaines et Sociales
- **SLSP** Statistical Language and Speech Processing
- **SLT** Speech Language Translation
- **SNR** Signal Noise Ratio
- **SPECOM** Speech Computer
- **SRAP** Système de Reconnaissance Automatique de la Parole
- **t-SNE** t-distributed Stochastic Neighbor Embedding : technique de réduction de dimension pour la visualisation de données
- **TA** Traduction Automatique
- **TAL** Traitement Automatique de la Langue
- **TALN** Traitement Automatique de la Langue Naturelle
- **TAP** Traduction Automatique de la Parole
- **TDNN** Time Delay Neural Network
- **TER** Translation Edit Rate
- **TERp** Translation Edit Rate plus
- **TIM** Textes, Informatique, Multilinguisme (équipe de recherche en traductologie à l'Université de Genève)
- **TRIDAN** Traduction et Reconnaissance d'Images de Documents Arabes Numérisés
- **UFSAC** Unification of Sense Annotated Corpora and Tools
- **UBM** Iniversal Background Model
- **VAD** Voice Activity Detection
- **WCE** Word Confidence Estimation
- **WER** Word Error Rate

- **Wordnet** WordNet® est une vaste base de données lexicales de l'anglais. Les noms, verbes, adjectifs et adverbes sont regroupés en ensembles de synonymes cognitifs (synsets), chacun exprimant un concept distinct. Les synsets sont reliés entre eux au moyen de relations conceptuelles-sémantiques et lexicales.
- **WSD** Word Sense Disambiguation

Annexe B

Digressions autour de la RAP

Encadrements : Solène Évain

Sélection de publications relatives à ce chapitre :

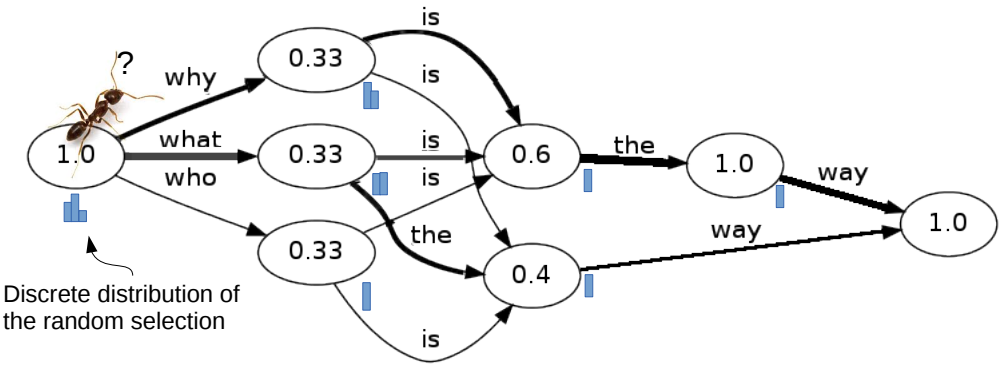
S. Evain, **B. Lecouteux**, D. Schwab, A. Contesse, A. Pinchaud, et al.. Human Beatbox Sound Recognition using an Automatic Speech Recognition Toolkit. Biomedical Signal Processing and Control 2021.

B. Lecouteux, D. Schwab. Ant colony algorithm applied to automatic speech recognition graph decoding. Interspeech 2015.

Dans cette annexe, j'aborde quelques travaux périphériques à mes axes de recherche. Je me suis notamment intéressé aux algorithmes d'expansion de graphe [Lecouteux and Schwab, 2014, Lecouteux and Schwab, 2015] et à la reconnaissance automatique du beatbox [Evain et al., 2019, Evain et al., 2020a, Evain et al., 2021a].

B.1 Décodage à partir d'un algorithme de fourmis

Dans [Lecouteux and Schwab, 2014, Lecouteux and Schwab, 2015] nous présentons un travail introduisant un nouveau paradigme d'exploration des graphes issus d'un système RAP. Une difficulté pour explorer un graphe de mots est d'appliquer sur ce dernier un modèle de langage d'ordre supérieur :



le nombre de chemins croît exponentiellement avec l'ordre du modèle de langage. Un graphe de mots décodé avec un modèle bi-gramme puis étendu avec un modèle 4-gramme peut multiplier sa taille par dix. Plusieurs techniques existent pour approcher (*compact-expansion*) ou effectuer des coupures dans la recherche (*beam-search*). Nous avons proposé une méthode alternative basée sur un algorithme constructif utilisé en recherche opérationnelle : les colonies de fourmis. Ce type d'algorithme a déjà été appliqué à d'autres problèmes en TAL, par exemple dans le cadre de la désambiguïsation lexicale [Schwab et al., 2013]. Nous avons proposé d'étendre ce paradigme à la RAP.

B.1.1 Algorithmes à colonies de fourmis

Les algorithmes se basant sur des colonies de fourmis sont des approches dites constructives : de nouvelles configurations sont générées en ajoutant itérativement des éléments de solution à la configurations en cours. Ils s'inspirent de la biologie et des observations portant sur le comportement social des fourmis. Ces insectes ont une capacité collective à trouver un plus court chemin entre une source d'énergie et leur fourmilière. Il a ainsi été démontré que la colonie s'auto-organise autour d'interactions entre individus autonomes. Ces interactions sont souvent simples et permettent à la colonie de résoudre des problèmes complexes. Ce phénomène est appelé intelligence en essaim [Bonabeau and Théraulaz, 2000].

B.1.2 Fourmis anamorphiques

Nous proposons plusieurs fourmis, comme dans [Bontoux, 2008] et qui utilisent des heuristiques pour biaiser l'exploration du graphe initial.

Des probabilités *a posteriori* sont calculées à partir du graphe bi-gramme extrait d'une première passe. Cette probabilité *a posteriori* est calculée via un

Données	Système de référence	Fourmi (3)	Fourmi (2)	Fourmi (1)	Fourmis (1)+(2)+(3)	Temps de Calcul référence/fourmis
Dev 2g	22.0%	22.0%	31.0%	29.3%	22.0%	1h/1h
Dev 3g	18.8%	20.8%	29.2%	26.4%	20.2%	13h/1h
Dev 4g	17.6%	19.5%	28.7%	26.0%	19.0%	18h/1h
Test 2g	21.6%	21.6%	30.1%	28.7%	21.6%	1h30/1h30
Test 3g	18.0%	20.0%	28.2%	25.9%	19.4%	19h/1h30
Test 4g	17.0%	19.1%	27.6%	25.0%	18.7%	27h/1h30

TABLE B.1 – Résultats, en terme de WER, des décodages avec l'algorithme à base de colonie de fourmis, les décodages présentent des oscillations de +/− 0.2 point d'une expérience à l'autre.

algorithme *forward-backward* comme présenté dans [Wessel et al., 2000].

Nous proposons plusieurs types de fourmis :

- Les fourmis à l'écoute (1) : elles sont influencées par la probabilité *a posteriori* acoustique du chemin suivant, sans considérer l'aspect linguistique.
- Les fourmis verbeuses (2) : elles sont uniquement influencées par la probabilité *a posteriori* linguistique du chemin suivant.
- Les fourmis oracles (3) : ce modèle hybride combine les deux précédents et correspond à la probabilité *a posteriori* du chemin suivant.

Ces différentes fourmis permettent d'orienter l'exploration en fonction d'informations différentes : linguistiques ou acoustiques.

Expériences préliminaires et résultats

La table B.1 présente les résultats (en terme de WER) obtenus avec notre algorithme. Orienter dès le départ les fourmis sur les meilleurs chemins influence beaucoup le résultat final. Les fourmis privilégiant l'acoustique ou la linguistique obtiennent de piètres résultats, mais combinées avec des fourmis hétérogènes, elles deviennent complémentaires.

Ces expériences montrent que l'algorithme de colonies de fourmis est capable d'étendre le graphe, sans toutefois atteindre les résultats *baseline*. Par contre, l'algorithme basé sur les fourmis est 15 fois plus rapide comme le montre la figure B.1 qu'une méthode d'extension classique.

Finalement, ce paradigme, basé sur les colonies de fourmis, permet de réaliser une exploration du graphe afin d'y trouver le chemin optimum sans

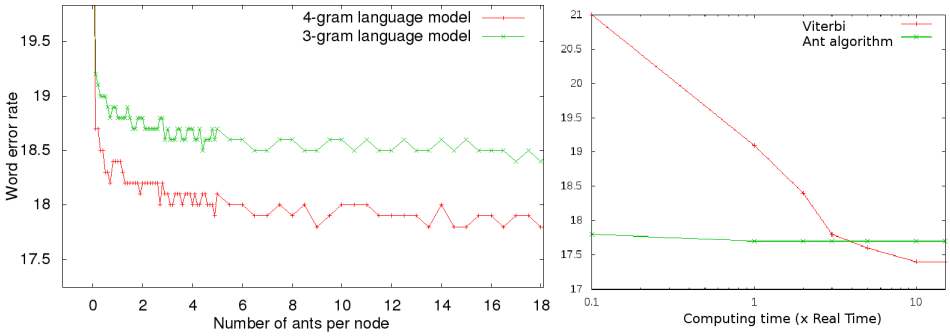


FIGURE B.1 – Temps de calcul comparés entre une expansion classique et l'algorithme à base de fourmis.

avoir à le re-construire dynamiquement. Les gains sont principalement liés à la vitesse d'exploration et la quantité de mémoire nécessaire. Ces travaux n'ont pas été poursuivis avec l'arrivée des systèmes de type bout-en-bout.

B.2 Reconnaissance automatique de parole beatboxée

Cette section aborde la reconnaissance automatique de la parole beatboxée. Ces travaux sont issus de la rencontre avec Nathalie Henrich-Bernadoni qui est Directrice de recherche CNRS au Gipsa-Lab (Grenoble). Elle s'intéresse notamment aux aspects articulatoires de la parole et lors de discussions relatives à l'annotation manuelle complexe de ses corpus, nous avons décidé de monter un projet visant l'aide à l'annotation [Evain et al., 2019].

Le Human Beatbox (en français, boîte à rythme humaine) est un art vocal urbain qui se développe depuis quarante ans. Il consiste à enchaîner des sons percussifs ou d'imitations instrumentales, avec une grande habileté articulatoire. Pratiqué par des amateurs ou des professionnels, il s'est développé avec les réseaux sociaux et l'organisation de championnats nationaux ou internationaux.

Cet art vocal a également attiré l'attention des cliniciens qui l'introduisent dans leur protocole de rééducation orthophonique des troubles de la voix et de la parole [Clouet and de Torcy, 2010, Bourdin and Navion, 2013, de Torcy et al., 2014] car il permet un travail approfondi et ludique de la coordination entre les gestes respiratoires, phonatoires et articulatoires [Paroni et al., 2019].

Dans ce contexte, le besoin d'un outil de reconnaissance automatique des sons beatboxés se fait grandissant. Dans le cadre des recherches en phonétique expérimentale ou clinique portant sur la production de sons beatboxés, l'annotation des bases de données se fait manuellement, impliquant un long travail qui pourrait être assisté par des annotations automatiques [Bourdin and Navion, 2013]. Quelques travaux ont exploré la reconnaissance automatique de beatbox[Sinyor et al., 2005, Picart et al., 2015], avec des sons isolés et sans considérer le beatbox comme un langage.

Nous présentons ci-après les travaux préliminaires que nous avons réalisés dans le cadre du stage de Master recherche de Solène Evain qui ont fait l'objet de plusieurs articles [Evain et al., 2019, Evain et al., 2020a, Evain et al., 2021a].

B.2.1 Constitution d'un corpus pour la reconnaissance du beatbox

Notre corpus de sons de beatbox a été enregistré par deux beatboxers : un professionnel et un amateur. Il est composé de 80 boxèmes (c'est le nom que nous avons donné aux unités de beatbox, en analogie avec les phonèmes) et peut être considéré comme un corpus de grand vocabulaire comparé aux précédents corpus utilisés dans les articles pour la classification de sons de beatbox.

Ce corpus a été enregistré avec six microphones. L'un d'entre eux était encapsulé (une ou deux mains couvrent la capsule du microphone). Les autres différaient en termes de spécificités (par exemple, condensateur ou dynamique) et de placement. La table B.2 donne les détails des microphones utilisés tandis que la table B.3 récapitule la composition du corpus.

B.2.2 Système RAP pour le beatbox

Notre hypothèse est que le beatbox peut se structurer à l'image d'un langage musical, utilisant les organes de la parole pour réaliser des sons qui peuvent être distingués les uns des autres et qui ont chacun une signification musicale spécifique. Dans ce contexte, un système de reconnaissance dédié à la parole peut permettre de reconnaître automatiquement les productions de beatbox. Ce travail préliminaire se concentre sur la reconnaissance de mots isolés. La co-articulation ou les limites des mots ont été écartées, tout en gardant des contraintes sur le traitement du bruit et la variabilité intra et inter-locuteur.

Type de Microphone	**Distance du microphone par rapport à la bouche**	**Spécificité du microphone**
Brauner VM1 (braun)	10 cm	condensateur + pop filter
DPA 4006 (ambia)	50 cm	condensateur micro ambiant
DPA 4060 (tie)	10 cm	condensateur
Shure SM58 (sm58p)	10 cm	dynamique
Shure SM58 (sm58l)	15 cm	dynamique
Shure beta 58 (beta)	1 cm	dynamique + encapsulé

Table B.2 – Spécifications des différents microphones

	Corpus de sons beatboxés
Beatboxers	Adrien (amateur), Andro (professionnel)
Date	2019
Nombre de sons différents	80
Nombre de sons différents prononcés par chaque beatboxer	Adrien : 56/80 Andro : 80/80
Microphone	5 + 1 encapsulé
Échantillonnage Précision	44100 Hz, 16 bits, mono, wav
Train	
Temps d'enregistrement	∼92mn
Nombre de répétitions	6 or 2
Test	
Temps d'enregistrement	∼114mn
Nombre de répétitions	7 (en moyenne)

Table B.3 – Récapitulatif du contenu de ce corpus

A : initiale, B: probabilité de silence 0,8 + pause,
C: B + 22 MFCC, D: B + 5 HMM

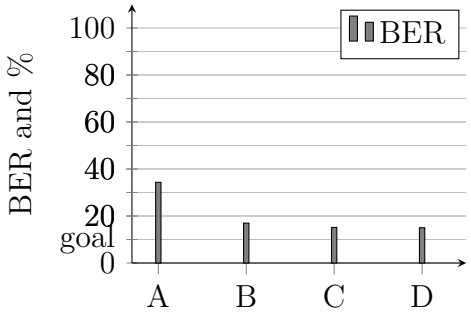


TABLE B.4 – Evolution du BER avec différentes configurations

Les traits acoustiques utilisés sont de type MFCC [Tiwari, 2010] et sont largement utilisées en RAP. Chaque son de beatbox a été associé à un HMM. Les détails du système sont exposés dans [Evain et al., 2019, Evain et al., 2020a]

Résultats

Plusieurs systèmes ont été entraînés dans le but de tester différents paramètres. L'influence de l'utilisation de différents microphones avec différents placements et sensibilités a également été étudiée afin de savoir si tous les enregistrements pouvaient être utilisés ensemble afin d'entraîner un système plus robuste. Les résultats avec nos différents systèmes (A : initial, B : probabilité de silence 0,8 + pause, C : B + 22 MFCC, D : B + 5 HMM) sont présentés dans la table B.5 et le graphe B.4.

Les principaux paramètres qui ont été testés sont : l'augmentation du nombre d'états HMM, l'augmentation du nombre de paramètres MFCC, l'ajout d'un phonème de pause dans le lexique et différentes probabilités de silence. Certains choix se sont basés sur l'article de [Picart et al., 2015].

Notre meilleur modèle a été obtenu avec une probabilité de silence de 0,8 , un phonème de silence "pause" ajouté dans les contextes droits et gauches et 22 paramètres MFCC. Le meilleur BER (Boxème Error Rate) obtenu a été de 15,13 %.

Nous avons pu observer que le type de microphone utilisé pour l'enregistrement avait peu d'influence sur le système. Son utilisation (encapsulé ou non) a par contre une incidence importante : le fait de ne pas prendre en compte le microphone encapsulé au sein des données d'entraînement s'est révélé bénéfique.

	A	B	C	D
Substitutions	19.19%	12.73%	10.70%	12.36%
Insertions	9.41%	0.18%	0.18%	0%
Deletion	5.72%	4.06%	4.24%	3.87%
CBR	75.09%	83.21%	85.06%	83.76%

TABLE B.5 – Insertions, substitutions, suppressions et Correct Boxeme Rate (CBR) pour les configurations A,B,C et D

Ce travail a permis de créer un premier système de reconnaissance de sons du beatbox et de produire un corpus pour l'entraînement et l'évaluation. Le projet dans lequel il s'est déroulé a été très intéressant du fait de la collaboration avec des artistes. Actuellement nous cherchons des financements pour lancer des recherches plus approfondies sur le sujet.

B.3 Conclusion

Dans ce chapitre nous avons présenté quelques travaux s'articulant autour de la reconnaissance automatique de la parole. Dans un premier temps, nous évaluons une nouvelle manière d'étendre les graphes de SRAP avec un algorithme à fourmis qui a montré des résultats encourageants. Puis nous présentons des travaux un peu plus récents qui portent sur la reconnaissance de sons *beatboxés* et les modèles auto-supervisés. Ces deux derniers travaux feront sans doute partie intégrante de mes activités de recherche dans les 5 prochaines années.

Annexe C

Liste des publications personnelles

Les publications sont consultables sur https://cv.archives-ouvertes.fr/benjamin-lecouteux.

C.1 Direction d'ouvrages (1)

[O1] L. Besacier, **B. Lecouteux**, G. Sérasset. Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1 : JEP. ATALA/AFCP, 2012.

C.2 Chapitres de livres avec comité de lecture (3)

[L1] M. Vacher, **B. Lecouteux**, F. Portet. On Distant Speech Recognition for Home Automation. Lecture Notes in Computer Science, vol.8700, Springer, p.161-188, février 2015, Smart Health : Open Problems and Future Challenges.

[L2] M. Vacher, **B. Lecouteux**, F. Portet PervasiveHealth Book, éditions Springer : 2014 Preliminary Experimental Results of Distant Speech Recognition for Home Automation.

[L3] M. Vacher, F. Portet, **B. Lecouteux**, C. Golanski. Telhealthcare Computing and Engineering : Principles and Design. Speech Analysis for Ambient Assisted Living : Technical and User Design of a Vocal Order System, (21) :607-638, CRC Press, Taylor and Francis Group, 2013.

C.3 Articles de journaux nationaux avec comité de lecture (2)

[j1] L. Vial, **B. Lecouteux**, D. Schwab. Approche supervisée à base de cellules LSTM bidirectionnelles pour la désambiguïsation lexicale. Traitement Automatique des Langues, ATALA, 2019.

[j2] Z. Elloumi, **B. Lecouteux**, O. Galibert, L. Besacier. Prédiction de performance des systèmes de reconnaissance automatique de la parole à l'aide de réseaux de neurones convolutifs. Traitement Automatique des Langues, ATALA, 2018.

C.4 Articles de journaux internationaux avec comité de lecture (11)

[j3] S. Evain, **B. Lecouteux**, D. Schwab, A. Contesse, A. Pinchaud, N. Henrich Bernardoni Human beatbox sound recognition using an automatic speech recognition toolkit. Biomedical Signal Processing and Control, 2021.

[j4] M. Vacher, F. Aman, S. Rossato, F. Portet, **B. Lecouteux**. Making emergency calls more accessible to older adults through a hands-free speech interface in the house. ACM Transactions on Accessible Computing , ACM New York, NY, 12 (2), pp.8 :1-8 :25, 2019.

[j5]N-T. Le, **B. Lecouteux**, L. Besacier. Automatic quality estimation for speech translation using joint ASR and MT features. Machine Translation, Springer Verlag, 2018.

[j6]**B. Lecouteux**, M. Vacher, F. Portet. Distant Speech Processing for Smart Home Comparison of ASR approaches in distributed microphone network for voice command. International Journal of Speech Technology, Springer Verlag 21, pp.601-618, 2018.

[j7]NQ. Luong, L. Besacier, **B. Lecouteux**. Find The Errors, Get The Better : Enhancing Machine Translation via Word Confidence Estimation. Natural Language Engineering, vol.23 n.4, Cambridge University Press, p.617-639, mars 2017.

[j8]NQ. Luong, L. Besacier, **B. Lecouteux**. Towards Accurate Predictors of Word Quality for Machine Translation : Lessons Learned on French-English and English-Spanish Systems. Data and Knowledge Engineering, vol.96-97, Elsevier, p.32-42, avril 2015.

[j9]M. Vacher, S. Caffiau, F. Portet, B. Meillon, C. Roux, E. Elias, **B. Lecouteux**, P. Chahuara. Evaluation of a context-aware voice interface for Ambient Assisted Living : qualitative user study vs. quantitative system evaluation. ACM Transactions on Accessible Computing, Special Issue on Speech and Language Processing for AT (Part 3), vol.7 issue 2, p.10-53, avril 2015.

[j10] NQ. Luong, L. Besacier, **B. Lecouteux**. Some Propositions to Improve the Prediction Capability of Word Confidence Estimation for Machine Translation. Journal of Computer Science and Communication Engineering, vol.30 n°3, p.36-49, août 2014.

[j11] **B. Lecouteux**, G. Linarès, Y. Estève, G. Gravier : Dynamic Combination of Automatic Speech Recognition Systems by Driven Decoding. » IEEE Transactions on Audio, Speech & Language Processing 21(6) : 1251-1260, 2013

[j12]**B. Lecouteux**, G. Linarès, S. Oger : Integrating imperfect transcripts into speech recognition systems for building high-quality corpora. Computer Speech & Language 26(2) : 67-89, 2012

[j13]M. Rouvier, G. Linarès, **B. Lecouteux** : Query-Driven Strategy for On-the-Fly Term Spotting in Spontaneous Speech. EURASIP J. Audio, Speech and Music Processing, 2010

C.5 Articles de Conférences et Workshops Internationaux avec comité de lecture (62)

[c0] S. Evain, H. Nguyen, H. Le, M. Zanon Boito, S. Mdhaffar, S. Alisamir, Z. Tong, N. Tomashenko, M. Dinarelli, T. Parcollet, A. Allauzen, Y. Estève, **B. Lecouteux**, F. Portet, S. Rossato, F. Ringeval, D. Schwab, L. Besacier. Task Agnostic and Task Specific Self-Supervised Learning from Speech with LeBenchmark. NeurIPS 2021 Track Datasets and Benchmarks. – en cours de soumission

[c1] S. Evain, H. Nguyen, H. Le, M. Zanon Boito, S. Mdhaffar, S. Alisamir, Z. Tong, N. Tomashenko, M. Dinarelli, T. Parcollet, A. Allauzen, Y. Estève, **B. Lecouteux**, F. Portet, S. Rossato, F. Ringeval, D. Schwab, L. Besacier. LeBenchmark : A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech. INTERSPEECH 2021.

[c2] H. Le, F. Barbier, H. Nguyen, N. Tomashenko, S. Mdhaffar, S. Gahbiche, F. Bougares , **B. Lecouteux**, D. Schwab, Y. Estève ON-TRAC' systems for the IWSLT 2021 low-resource speech translation and multilingual speech translation shared tasks. IWSLT.

[c3] L. Ormachea Grijalba, P. Bouillon, J. Gerlach, **B. Lecouteux**, D. Schwab, H. Spechbach. Reconnaissance vocale du discours spontané pour le domaine médical. Conférence sur les Technologies du Langage Humain, 2021.

[c4]D. Schwab, P. Trial, C. Vaschalde, L. Vial, E. Esperança-Rodier, **B. Lecouteux**. Providing semantic knowledge to a set of pictograms for people with disabilities : a set of links between WordNet and Arasaac : Arasaac-WN. LREC, 2020, Marseille, France.

[c5]S. Evain, **B. Lecouteux**, F. Portet, I. Estève, M. Fabre. Towards Automatic Captioning of University Lectures for French Students who are Deaf (papier court). ACM SIGACCESS Conference on Computers and Accessibility, 2020, Athènes, Greece.

[c6]L. Chasseur, M. Dohen, **B. Lecouteux**, S. Riou, Am. Rochet-Capellan, D. Schwab.. Evaluation of the acceptability and usability of augmentative and alternative communication (AAC) tools : the example of pictogram grid communication systems with voice output (papier court). ACM SIGACCESS Conference on Computers and Accessibility, 2020, Athènes, Greece.

[c7]M. Elbayad, H. Nguyen, F. Bougares, N. Tomashenko, A. Caubrière, **B. Lecouteux**, Y. Estève, L. Besacier. ON-TRAC Consortium for End-to-End and Simultaneous Speech Translation Challenge Tasks at IWSLT 2020. The International Conference on Spoken Language Translation ACL - 17th IWSLT, Jul 2020, Seattle, WA, United States.

[c8]H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, **B. Lecou-**

teux, A. Allauzen, B. Crabbé, L. Besacier, D. Schwab. . FlauBERT : Unsupervised Language Model Pre-training for French. LREC, 2020, Marseille, France.

[c9]L. Vial, **B. Lecouteux**, D. Schwab, H. Le, L. Besacier. The LIG system for the English-Czech Text Translation Task of IWSLT 2019. IWSLT (16th International Workshop on Spoken Language Translation), 2019, Hong-Kong, China.

[c10]L. Vial, **B. Lecouteux**, D. Schwab. Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. Global Wordnet Conference, 2019, Wroclaw, Poland.

[c11] F. Portet, S. Caffiau, F. Ringeval, M. Vacher, N. Bonnefond, S. Rossato, **B. Lecouteux**, T. Desot. Context-Aware Voice-based Interaction in Smart Home -VocADom@A4H Corpus Collection and Empirical Assessment of its Usefulness. PICom 2019 - 17th IEEE International Conference on Pervasive Intelligence and Computing, Aug 2019, Fukuoka, Japan. pp.811–818,

[c12] S. Evain, A. Contesse, A. Pinchaud, D. Schwab, **B. Lecouteux**, N. Henrich. Beatbox sounds recognition using a speech-dedicated HMM-GMM based system. Models and Analysis of Vocal Emissions for Biomedical Applications : 11th International Workshop, Dec 2019, Firenze, Italy.

[c13]Z. Elloumi, L. Besacier, O. Galibert, J. Kahn, **B. Lecouteux**. ASR performance prediction on unseen broadcast programs using convolutional neurol networks. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr 2018, Calgary, Alberta, Canada.

[c14]L. Vial, **B. Lecouteux**, D. Schwab. UFSAC : Unification of Sense Annotated Corpora and Tools. Language Resources and Evaluation Conference (LREC), May 2018, Miyazaki, Japan.

[c15]Z. Elloumi, L. Besacier, O. Galibert, **B. Lecouteux**. Analyzing Learned Representations of a Deep ASR Performance Prediction Model. Blackbox NLP Workshop at EMNLP 2018, Nov 2018, Bruxelles, Belgium.

[c16] C. Vaschalde, P. Trial, E. Esperança-Rodier, D. Schwab, **B. Lecouteux**. Automatic pictogram generation from speech to help the implementation of a mediated communication. Conference on Barrier-free

Communication, Nov 2018, Geneva, Switzerland.

[c17]NT. Le, **B. Lecouteux**, L. Besacier. Disentangling ASR and MT Errors in Speech Translation. Machine Translation Summit, septembre 2017, Nagoya, Japon, p.312-323

[c18] L. Vial, **B. Lecouteux**, D. Schwab. Sense Embeddings in Knowledge-Based Word Sense Disambiguation. International Conference on Computational Semantics, p.W17-6940, Septembre 2017, Montpellier, France

[c19]M. Vacher, S. Bouakaz, ME. Bobillier-Chaumon, F. Aman, RA. Khan, S. Bekkadja, F. Portet, E. Guillou, S. Rossato, **B. Lecouteux**. The CIRDO Corpus : Comprehensive Audio/Video Database of Domestic Falls of Elderly People.International Conference on Language Resources and Evaluation (LREC), Portoroz, Slovenie. p.1389-1396, mai 2016

[c20]F. Aman, M. Vacher, F. Portet, W. Duclot, **B. Lecouteux**. CirdoX : an On/Off-line Multisource Speech and Sound Analysis Software. Language Resources and Evaluation Conference (LREC), Portoroz, Slovenie. p.1978-1985, mai 2016

[c21]NT. Le, C. Servan, **B. Lecouteux**, L. Besacier. Better Evaluation of ASR in Speech Translation Context Using Word Embeddings. InterSpeech, San-Francisco, USA, p.2538-2542, septembre 2016

[c22]NT. Le, **B. Lecouteux**, L. Besacier. Joint ASR and MT Features for Quality Estimation in Spoken Language Translation. International Workshop on Spoken Language Translation (IWSLT), Seattle, USA, papier num 13, décembre 2016

[c23] M. Vacher, **B. Lecouteux**, F. Aman, S. Rossato, F. Portet. Recognition of Distress Calls in Distant Speech Setting : a Preliminary Experiment in a Smart Home. Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), Dresde, Allemagne, p.1-7, septembre 2015

[c24] L. Besacier, **B. Lecouteux**, NQ. Luong, NT. Le. Spoken language translation graphs re-decoding using automatic quality assessment. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Scotsdale, Arizona, USA, p.267-274, décembre 2015

[c25] C. Servan, NT. Le, NQ. Luong, **B. Lecouteux**, L. Besacier. An Open Source Toolkit for Word-level Confidence Estimation in Machine Translation. International Workshop on Spoken Language Translation (IWSLT), Da Nang, Vietname, p.196-203, décembre 2015

[c26] S. Samson, L. Besacier, **B. Lecouteux**, M. Dyab. Using Resources from a Closely-related Language to Develop ASR for a Very Under-resourced Language : A Case Study for Iban. Interspeech, Dresde, Allemagne, p.1270-1274, septembre 2015

[c27] M. Vacher, **B. Lecouteux**, J. Serrano-Romero, M. Ajili, F. Portet, et al.. Speech and Speaker Recognition for Home Automation : Preliminary Results. IEEE International Conference Speech Technology and Human-Computer Dialogue (SPED), Bucarest, Roumanie, p.181-190, octobre 2015

[c28] S. Samson Juan, L. Besacier, **B. Lecouteux**, TP. Tan. Merging of Native and Non-native Speech for Low-resource Accented ASR. International Conference on Statistical Language and Speech Processing (SLSP), Budapest, Hongrie, p.255-266, novembre 2015

[c29]**B. Lecouteux**, D. Schwab. Ant colony algorithm applied to automatic speech recognition graph decoding. Interspeech, Dresde, Allemagne, p.2122-2126, septembre 2015

[c30] S. Samson Juan, L. Besacier, **B. Lecouteux**, TP. Tan. Using closely-related language to build an ASR for a very under-resourced language : Iban. IEEE Oriental International Committee for Co-ordination and Standardisation of Speech Databases (O-COCOSDA), Phuket, Thaïlande, p.1-5, septembre 2014

[c31]M. Vacher, **B. Lecouteux**, F. Portet. Multichannel Automatic Recognition of Voice Command in a Multi-Room Smart Home : an Experiment involving Seniors and Users with Visual Impairment. Interspeech, Singapour, septembre 2014, p.1008-1012

[c32] NQ. Luong, L. Besacier, **B. Lecouteux**. An Efficient Two-Pass Decoder for SMT Using Word Confidence Estimation. European Association for Machine Translation (EAMT), Dubrovnik, Croatie, p.117-124, juin 2014

[c33] TP. Tan, L. Besacier, **B. Lecouteux**. Acoustic Model Merging

Using Acoustic Models from Multilingual Speakers for Automatic Speech Recognition. International Conference on Asian Language Processing (IALP), Sarawak, Malaisie, p.1-4, octobre 2014

[c34] M. Vacher, **B. Lecouteux**, P. Chahuara, F. Portet, B. Meillon, et al.. The Sweet-Home speech and multimodal corpus for home automation interaction. Language Resources and Evaluation Conference (LREC), Reykjavik, Islande, p.4499-4506, mai 2014.

[c35] NQ. Luong, L. Besacier, **B. Lecouteux**. LIG System for Word Level QE task at WMT14. Workshop on Machine Translation (WMT), p.335-341, Baltimore, USA, juin 2014.

[c36]L. Besacier, **B. Lecouteux**, NQ. Luong, K. Hour, M. Hadjsalah. Word confidence estimation for speech translation. International Workshop on Spoken Language Translation (IWSLT),South Lake Tahoe, USA, p.169-175, décembre 2014

[c37] NQ. Luong, L. Besacier, **B. Lecouteux**. Word Confidence Estimation for SMT N-best List Re-ranking. Proceedings of the Workshop on Humans and Computer-assisted Translation (HaCaT) during EACL, Gothenburg, Suède, p.1-9, avril 2014

[c38]N.Q. Luong, **B. Lecouteux**, L. Besacier. LIG System for WMT13 QE Task : Investigating the Usefulness of Features in Word Confidence Estimation for MT. Workshop on Statistical Machine Translation, Sofia, Bulgaria, aug 2013.

[c39]N.Q. Luong, L. Besacier, **B. Lecouteux**. Word Confidence Estimation and its Integration in Sentence Quality Estimation for Machine Translation. In Proceedings of the fifth international conference on knowledge and systems engineering (KSE), Hanoi, Vietnam, oct 2013.

[c40] M. Vacher, **B. Lecouteux**, D. Istrate, T. Joubert, F. Portet, P. Chahuara (2013) : Experimental Evaluation of Speech Recognition Technologies for Voice-based Home Automation Control in a Smart Home SLPAT 2013

[c41]M. Vacher, **B. Lecouteux**, D. Istrate, T. Joubert, F. Portet, M. A. Sehili, P. Chahuara : Evaluation of a real-time voice order

recognition system from multiple audio channels in a home. INTERSPEECH 2013 : 2062-2064, 2013

[c42]M. Vacher, P. Chahuara, **B Lecouteux**, D. Istrate, F. Portet, T. Joubert (2013) : The Sweet-Home Project : Audio Processing and Decision Making in Smart Home to Improve Well-being and Reliance Engineering in Medicine and Biology Society, EMBC, 2013 Annual International Conference of the IEEE, 2013

[c43] M. A. Sehili, **B. Lecouteux**, M. Vacher, F. Portet, D. Istrate, B. Dorizzi, J. Boudy (2012) : Sound Environment Analysis in Smart Home. Ambiant Intelligence (AMI) 2012 : 208-223

[c44]L. Besacier, **B. Lecouteux**, M. Azouzi, N.Q. Luong (2012). The LIG English to French Machine Translation System for IWSLT 2012. In proceedings of the 9th International Workshop on Spoken Language Translation (IWSLT), Hong Kong.

[c45]M. Vacher, **B. Lecouteux**, F. Portet (2012) : Recognition of Voice Commands by Multisource ASR and Noise Cancellation in a Smart Home Environment. EUSIPCO (European Signal Processing Conference), :1663-1667, Bucarest, Romania, aug 2012.

[c46] **B. Lecouteux**, M. Vacher, F. Portet (2011) : Distant Speech Recognition for Home Automation : Preliminary Experimental Results in a Smart Home. IEEE SPED 2011, :41-50, Brasow, Romania, may 2011.

[c47]**B. Lecouteux**, L. Besacier, H. Blanchon (2011) : LIG English-French spoken language translation system for IWSLT 2011. IWSLT 2011 : 68-72

[c48]G. Senay, G. Linarès, **B. Lecouteux** (2011) « A segment-level confidence measure for spoken document retrieval » *In Proceedings of the International Conference on Acoustic Speech and Signal Processing, ICASSP'11*, Prague, République Tchèque .

[c49]**B. Lecouteux**, M. Vacher, F. Portet (2011) « Distant Speech Recognition in a Smart Home : Comparison of Several Multisource ASRs in Realistic Conditions. » *International conference of the Speech Communication Association, ISCA*.InterSpeech'11.

[c50]M. Vacher, D. Istrate, F. Portet, T. Joubert, T. Chevalier,

S. Smidtas, B. Meillon, **B. Lecouteux**, M. Sehili, P. Chahuara and S. Méniard (2011) « The Sweet-Home project : audio technology in smart homes to improve wellbeing and reliance. » *IEEE Engineering in Medicine and Biology Society*, EMBC'11

[c51]M. Potet, R. Rubino, **B. Lecouteux**, S. Huet, H. Blanchon, L. Besacier, F. Lefevre « The LIG/LIA Machine Translation System for WMT 2011 » *Conference on Empirical Methods in Natural Language Processing, EMNLP (2011).*

[c52] G. Senay, G. Linarès, **B. Lecouteux**, S. Oger, T. Michel (2010). « Transcriber driving strategies for transcription aid system » *In Language Resources and Evaluation Conference, LREC'10*, Malta.

[c53]**B. Lecouteux**, R. Rubino, G. Linarès(2010) « Improving backoff models with bag of words and hollow-grams » *International conference of the Speech Communication Association, ISCA*, InterSpeech'10, Tokyo, JP.

[c54] **B. Lecouteux**, P. Nocera, G. Linarès (2010). « Semantic cache model driven speech recognition » *In Proceedings of the International Conference on Acoustic Speech and Signal Processing, ICASSP'10*, Dallas, USA.

[c55]B. Lecouteux, G. Linarès, B. Favre. Combined low level and high level features for Out-Of-Vocabulary Word detection. Interspeech 2009, 2009, Brighton, United Kingdom.

[c56]M. Rouvier, G. Linarès, B. Lecouteux. On-the-fly term spotting by phonetic filtering and request-driven decoding. IEEE Spoken Language Technology Workshop (SLT), Dec 2008, Goa, India.

[c57] **B. Lecouteux**, G. Linarès. Using prompts to produce quality corpus for training automatic speech recognition systems. MELECON 2008 - The 14th IEEE Mediterranean Electrotechnical Conference , May 2008, Ajaccio, France.

[c58] **B. Lecouteux**, G. Linarès, Y. Estève, G. Gravier. Generalized Driven Decoding for Speech Recognition System Combination. IEEE International Conference on Acoustics, Speech and Signal Processing, Mar 2008, Las Vegas, United States.

[c59]**B. Lecouteux**, G. Linarès, F. Beaugendre, P. Nocera. Text

island spotting in large speech databases. INTERSPEECH, Aug 2007, Anvers, Belgium.

[c60]**B. Lecouteux**, G. Linarès, Y. Estève, J. Mauclair. System Combination by Driven Decoding. IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07 , Apr 2007, Honolulu, United States.

[c61]**B. Lecouteux**, G. Linarès, P. Nocera, J-F. Bonastre. Imperfect transcript driven speech recognition. INTERSPEECH, Sep 2006, Pittsburgh, United States.

C.6 Articles de conférences nationales avec comité de lecture (24)

[c62]H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, **B. Lecouteux**, A. Allauzen, B. Crabbé, L. Besacier, D. Schwab. FlauBERT : des modèles de langue contextualisés pré-entraînés pour le français. Volume 2 : Traitement Automatique des Langues Naturelles (TALN), Jun 2020, Nancy, France. pp.268-278.

[c63] S. Evain, A. Contesse, A. Pinchaud, D. Schwab, **B. Lecouteux**, N. Henrich. Reconnaissance de parole beatboxée à l'aide d'un système HMM-GMM inspiré de la reconnaissance automatique de la parole. 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), 2020, Nancy, France. pp.208-216.

[c64] D. Schwab, P. Trial, V. Céline, L. Vial, **B. Lecouteux**. Apporter des connaissances sémantiques à un jeu de pictogrammes destiné à des personnes en situation de handicap : Un ensemble de liens entre Princeton WordNet et Arasaac, Arasaac-WN. 26e Conférence sur le Traitement Automatique des Langues Naturelles, 2019, Toulouse, France. pp.619-622.

[c65]Best Paper L. Vial, **B. Lecouteux**, D. Schwab. Compression de vocabulaire de sens grâce aux relations sémantiques pour la désambiguïsation lexicale. TALN 2019 (Conférence sur le Traitement Automatique des Langues Naturelles), Jul 2019, Toulouse, France.

[c66]D. Schwab, P. Trial, C. Vaschalde, L. Vial, **B. Lecouteux**. Apporter des connaissances sémantiques à un jeu de pictogrammes destiné à

des personnes en situation de handicap : Un ensemble de liens entre Wordnet et Arasaac, Arasaac-WN. TALN 2019, 2019, Toulouse, France.

[c67] L. Vial, **B. Lecouteux**, D. Schwab. Approche supervisée à base de cellules LSTM bidirectionnelles pour la désambiguïsation lexicale. 25e conférence sur le Traitement Automatique des Langues Naturelles, May 2018, Rennes, France.

[c68] M. Hadj Salah, L. Vial, H. Blanchon, M. Zrigui, **B. Lecouteux**, D. Schwab. La désambiguïsation lexicale d'une langue moins bien dotée, l'exemple de l'arabe. 25e conférence sur le Traitement Automatique des Langues Naturelles, May 2018, Rennes, France.

[c69] C. Vaschalde, **B. Lecouteux**, D. Schwab. Génération de pictogrammes à partir de la parole spontanée pour la mise en place d'une communication médiée. 50 ans de linguistique sur corpus oraux : Apports à l'étude de la variation, Nov 2018, Orléans, France.

[c70] L. Vial, **B. Lecouteux**, D. Schwab. Représentation vectorielle de sens pour la désambiguïsation lexicale à base de connaissances. Conférence sur le Traitement Automatique des Langues Naturelles (TALN), papier court, Orléans, vol.2 p.142-149, juin 2017

[c71] L. Vial, **B. Lecouteux**, D. Schwab. Uniformisation de corpus anglais annotés en sens. Traitement Automatique des Langues Naturelles (TALN), démonstration, Orléans, vol.3 p.27-30, juin 2017

[c72] K. Bouzidi, Z. Elloumi, L. Besacier, **B. Lecouteux**, MF. Benzeghiba. Traitement des Mots Hors Vocabulaire pour la Traduction Automatique de Document OCRisés en Arabe. Traitement Automatique des Langues Naturelles (TALN), papier long, Orléans,vol.1 p.63-76 , juin 2017

[c73] M. Vacher, **B. Lecouteux**, F. Aman, F. Portet, S. Rossato. Acquisition et reconnaissance automatique d'expressions et d'appels vocaux dans un habitat. Journées d'Étude de la parole (JEP), Paris, p.28-36, juillet 2016

[c74] L. Besacier, **B. Lecouteux**, NQ. Luong. Utilisation de mesures de confiance pour améliorer le décodage en traduction de parole. Traitement Automatique des Langues Naturelles (TALN), papier long, Caen, p.244-254, juin 2015

[c75] M. Vacher, F. Portet, F. Aman, **B. Lecouteux**, S. Rossato, et al.. Reconnaissance automatique de la parole dans les habitats intelligents : application à l'assistance à domicile (SFTAG). Journées Annuelles de la Société Française des Technologies pour l'Autonomie et de Gérontechnologie, p.38-41, Paris, novembre 2014

[c76] **B. Lecouteux**, D. Schwab. Décodage de graphe à l'aide de colonies de fourmis. Journées d'Étude de la Parole (JEP), Le Mans, p.302-310, juin 2014

[c77] **B. Lecouteux**, L. Besacier (2013) : Vers un décodage guidé pour la traduction automatique, TALN 2013

[c78] **B. Lecouteux**, M. Vacher, F. Portet (2012) : Reconnaissance d'ordres domotiques en conditions bruitées pour l'assistance à domicile. JEP-TALN-RECITAL 2012, Atelier ILADI 2012 :31–39, Grenoble, France, juin 2012

[c79] **B. Lecouteux**, M. Vacher, F. Portet (2012) : Reconnaissance automatique de la parole distante dans un habitat intelligent : méthodes multi-sources en conditions réalistes JEP-TALN-RECITAL 2012, volume 1 : JEP, Grenoble, France, juin 2012

[c80] G. Senay, **B. Lecouteux**, G. Linarès (2012) : Prédiction de l'indexabilité d'une transcription-Prediction of transcription indexability JEP-TALN-RECITAL 2012, volume 1 : JEP, Grenoble, France, juin 2012

[c81] **B. Lecouteux**, P. Nocera, G. Linarès (2010). « Décodage guidé par un modèle cache sémantique ». *Journées d'études sur la parole, JEP'2010*, Mons, Belgique.

[c82] G. Senay, G. Linarès, **B. Lecouteux**, S. Oger, T. Michel (2010). « Décodage intéractif de la parole ». *Journées d'études sur la parole, JEP'2010*, Mons, Belgique.

[c83] **B. Lecouteux**, G. Linarès, B. Favre. Détection de mots hors-vocabulaire par combinaison de mesures de confiance de haut et bas niveaux. MajecSTIC 2009 , Nov 2009, Avignon, France.

[c84] **B. Lecouteux**, G. Linarès, Y. Estève, G. Gravier. Combinai-

son de systèmes par décodage guidé. JEP / TALN / RECITAL 2008, Jun 2008, Avignon, France.

[c85] **B. Lecouteux**, G. Linarès, P. Nocera, J-F. Bonastre. Reconnaissance de la parole guidée par des transcriptions approchées. Journées d'Etudes sur la Parole (JEP), Jun 2006, Dinard, France.

C.7 Rapports (11)

[r1] L. Ormaechea Grijalba, J. Gerlach, D. Schwab, P. Bouillon, **B. Lecouteux** Building and enhancement of an ASR system for emergency medical settings : towards a better accessibility for allophone and disabled patients. [Research Report] LIG. 2020.

[r2] Z. Elloumi, O. Galibert, **B. Lecouteux**, L. Besacier. Investigating robustness of a deep ASR performance prediction system. [Research Report] LIG lab ; LNE. 2019.

[r3] L. Vial, **B. Lecouteux**, D. Schwab. Approche supervisée à base de cellules LSTM bidirectionnelles pour la désambiguïsation lexicale.. [Rapport de recherche] UGA - Université Grenoble Alpes. 2018.

[r4] L. Vial, **B. Lecouteux**, D. Schwab. WSD. [Research Report] LIG. 2018.

[r5] C. Vaschalde, P. Trial, E. Esperança-Rodier, **B. Lecouteux**, D. Schwab. Automatic pictogram generation from speech to help the implementation of a mediated communication. [Research Report] LIG ; UGA (Université Grenoble Alpes). 2018.

[r6] L. Vial, **B. Lecouteux**, D. Schwab. UFSAC : Unification of Sense Annotated Corpora and Tools. [Research Report] UGA - Université Grenoble Alpes. 2017.

[r7] R. Rubino, **B. Lecouteux**, G. Linarès. Amélioration des modèles de repli par des sacs de mots et des n-grammes à variables. [Rapport de recherche] LIG. 2016.

[r8] S. Khandelwal, **B. Lecouteux**, L. Besacier. Comparing GRU and LSTM for automatic speech recognition. [Research Report] LIG. 2016.

[r9] **B. Lecouteux**, L. Besacier. Use of auxiliary translation for improving decoding in statistical machine translation. [Research Report] LIG. 2016.

[r10] G. Linarès, **B. Lecouteux**, D. Matrouf, P. Nocera Phone duration models for fast broadcast news transcriptions. [Research Report] LIA. 2006.

[r11] **B. Lecouteux**. Alignement de transcriptions imparfaites sur un flux de parole. [Rapport de recherche] LIA. 2005.