CFA/VISHNO 2016

Adaptation(s) d'un Système de Sous-titrages en faible latence

C. Lailler, Y. Estève et P. Deléglise
LIUM, Université du Maine, avenue Laennec, 72085 Le Mans, France
carole.lailler@lium.univ-lemans.fr

Les travaux présentés portent sur une analyse comparative de sorties d'un Système de Reconnaissance Automatique (dorénavant ASR) de la Parole. L'objectif est de mesurer non seulement la robustesse du système utilisé mais également les diverses possibilités pour faciliter une adaptation acoustique, phonétique et phonologique afin d'être le plus fidèle possible aux énoncés proférés en parole spontanée. Nous nous proposons d'étudier deux sorties issues du Système ASR à faible latence développé au LIUM, l'une provenant d'un système très généraliste, l'autre d'un système adapté à dessein. L'objectif de cette présentation sera double : il s'agira non seulement de présenter le système à latence faible et l'évolution de son architecture mais aussi de discuter, au sein d'une analyse comparative, de la nécessité d'une adapation dans un cadre applicatif très spécialisé. Nous aborderons également la question des coûts tant humains que techniques et nous rendrons compte des résultats obtenus.

1 Introduction

Le présent article a pour but de présenter une analyse comparative de sorties du Système de Reconnaissance Automatique (dorénavant ASR) de la Parole, à latence faible, développé au Laboratoire d'Informatique de l'Université du Maine (LIUM). L'objectif est non seulement d'éprouver la robustesse du système utilisé mais également de mesurer les effets d'une adaptation acoustique, phonétique et phonologique lorsque le besoin s'en fait sentir : adaptation pour des réunions de spécialistes (chercheurs, professionnels de santé, industriels…), adaptation pour des présentations en public avec sous-titrages à destination des mal-entendants constituent autant d'exemples possibles. Plus la parole est spontanée, c'est-à-dire sans préparation préalable de la part du locuteur, plus il peut être difficile pour un système ASR de transcrire les propos au plus près de la réalité. Toutefois, à cette première difficulté s'ajoute une seconde liée au vocabulaire de spécialité et aux tournures de phrases employées. La problématique de nos travaux est alors de conserver des résultats réellement corrects, mesurables en Word Error Rate (dorénavant WER) satisfaisants, lorsque le discours oral fait usage d'une terminologie idoine au domaine évoqué.

Pour être efficace un système ASR doit être doté d'un Modèle de Langage et d'un Modèle Acoustique performants. Ces modèles doivent faire l'objet d'un apprentissage sur des données nombreuses et pondérées pour qu'ils puissent être en mesure de faire état de la langue parlée. Un système doit donc être suffisamment générique pour couvrir le plus grand nombre d'aspects langagiers usuels. Cependant, il doit également permettre de reconstituer la totalité du message, y compris lorsque celui-ci est bruité et/ou contient un vocabulaire spécifique. Si l'on comprend aisément que les noms propres (noms de scientifiques, de lieux, de théorèmes, de principes, de matériel, etc.) peuvent constituer des freins pour le travail de transcriptions automatiques, les lexies issues d'un vocabulaire terminologique constituent de véritables sources d'erreurs. Nous nous proposons ici de comparer des sorties issues du Système ASR développé au LIUM. Les premières sorties se fondent sur le système dit "classique", construit avec un modèle de langage probabiliste générique, sans adaptation particulière et destiné à la transcription de discours officiels. Les secondes sont issues du même système, après diverses adaptations, pour permettre le sous-titrage en français de la soutenance de thèse d'une doctorante en acoustique. L'objectif de cette présentation sera double : il s'agira non seulement de présenter le système et l'évolution de son architecture mais aussi de discuter, au sein d'une analyse comparative dans un cadre applicatif très spécialisé, des adaptations mises en place.

Nous étudierons, dans un premier temps, l'architecture du système utilisé et évoquerons les problèmes inhérents à ce type de travaux. Puis, nous nous intéresserons aux voies d'adaptations proposées pour permettre le sous-titrage d'une thèse en acoustique avant de discuter plus en détails les résultats obtenus ainsi que les nouvelles possibilités d'adaptation envisagées. Quelques éléments de conclusion et de perspective termineront cet article.

2 Vers un système ASR à latence faible

Un système ASR, à l'heure actuelle, est fondé sur un ensemble de méthodes statistiques, qui peuvent légèrement différées selon les besoins (au plus proche du temps réel ou non) et selon les possibilités (matériel à disposition).

2.1 Architecture et Fonctionnement d'un système ASR

Pour fonctionner, un système ASR a besoin de 3 bases de connaissances :

- Tout d'abord, il faut un ensemble de modèles de markov cachés (MMC). Cet ensemble modélise l'acoustique des différents phonèmes d'une langue et est appris sur un ensemble de corpus de parole associés à une transcription manuelle, ou plus simplement à des sous-titres.
- Un vocabulaire ainsi qu'un modèle de langage sont également nécessaires. Le modèle de langage résulte d'un apprentissage sur un ensemble de textes qui permet d'associer une probabilité à toute suite de mots du vocabulaire.
- Enfin, un dictionnaire de phonétisation permet, quant à lui, d'associer à chaque mot du vocabulaire la suite des modèles acoustiques représentant le mot. Ce dictionnaire de phonétisation est issu de systèmes automatiques et/ou de systèmes à base de règles et peut, éventuellement, faire l'objet de vérifications manuelles partielles.

Cet ensemble de modèles va permettre de trouver la séquence de mots W_h qui vérifie (1).

$$W_h = \underset{W}{\operatorname{argmax}}\, P(W|x) \tag{1}$$

En effet en utilisant la formule de Bayes (2) et en constatant que $P(x)$ est constant pour toutes les séquences de mots. L'équation (1) devient (3).

$$P(W|x) = \frac{P(x|W)P(W)}{P(x)} \quad (2)$$

$$W_h = \underset{W}{\operatorname{argmax}}\, P(x|W)P(W) \quad (3)$$

Pour comparer effectivement les probabilités de deux hypothèses, la valeur de $P(x|W)$ est calculée par le modèle acoustique tandis que la valeur de $P(W)$ est calculée par le modèle de langage. Comme il n'est pas possible d'énumérer toutes les séquences possibles, on utilise un algorithme dont la complexité provient de :

- l'ordre du modèle de langage. Ainsi, pour un modèle d'ordre 3, il faut connaître les 2 mots précédents afin de calculer la probabilité du troisième. Plus l'ordre est grand, plus le modèle de langage permet de lever les ambiguïtés laissées par le modèle acoustique. A contrario, plus cet ordre est élevé, plus la complexité du décodage est importante. Elle oblige alors à effectuer des élagages importants qui sont susceptibles de nuire à la qualité du décodage.

- La complexité est également due à l'utilisation de modèles contextuels pour l'acoustique. Ces derniers entraînent une dépendance du modèle courant par rapport au suivant, particulièrement lorsqu'on se trouve à la jonction entre deux mots.

2.2 Vers un décodage en temps réel

Afin de permettre une transcription *en live* (lors d'un sous-titrage par exemple), il est important d'avoir des systèmes dont la sortie est livrée presque immédiatement après le flux de paroles prononcé. C'est la raison pour laquelle il convient de s'intéresser à la problématique de la rapidité des systèmes. Pour ce faire, on examinera en détails la modélisation acoustique, son évolution et les algorithmes de décodage.

2.2.1 Modélisation acoustique

La modélisation acoustique constitue un incontournable de la construction d'un système ASR. Jusqu'à récemment, les systèmes utilisaient une modélisation des états des MMC par des mélanges de gaussiennes (GMM). La dimension de ces gaussiennes étaient de l'ordre de 40, ce qui correspondaient à une paramétrisation utilisant 13 coefficients MFFC (Mel Frequency Cepstral Coefficients) auxquels on ajoutait leur dérivée et leur accélaration discrètes. Ils étaient calculés sur une fenêtre glissante de 25 ms.

Pour obtenir un système relativement robuste, l'utilisation des GMM nécessite une adaptation aux variations des conditions acoustiques. Cette dernière s'effectue par le calcul d'une transformation linéaire des paramètres calculée pour des zones de signal homogène, tant en locuteur qu'en conditions acoustiques. Pour cela, il faut, préalablement au décodage, découper le signal en zones homogènes. Deux passes de décodage sont également nécessaires : la première a pour but de calculer la transformation, la deuxième l'utilise. Par ailleurs, il faut noter que la présence d'un bruit dans une bande de fréquence perturbe l'ensemble des coefficients.

Récemment l'apparition des réseaux de neurones profonds [1], c'est à dire possédant au moins 5 ou 6 couches cachées, a permis de modéliser directement la probabilité acoustique d'un état d'un HMM. L'utilisation de ces Deep Neural Networks (DNN) a également autorisé l'augmentation de la dimension des vecteurs de paramètres allant jusqu'à des tailles de vecteurs de plusieurs centaines. On a ainsi pu envisager une modélisation des trames acoustiques par les TRAPs (Temp oRAl Patterns). Le principe de ces TRAPS est d'effectuer une transformation en cosinus sur la sortie de 23 bancs filtres répartis suivant un échelle MEL. Cette transformation est calculée sur une durée de 150 ms, la valeur de chaque filtre étant calculée toutes les 10 ms sur une fenêtre de 25ms.

Ce couplage DNN/TRAP a généré des avancées importantes en matière de modélisation acoustique. Ainsi, on a pu mieux prendre en compte la variation du spectre par l'empan des paramètres. Or, l'on sait parfaitement, qu'en matière de parole, les variations importent davantage que la forme même. Grâce à ce diptyque DNN/TRAP, on a non seulement pu limiter l'influence du bruit à sa bande de fréquence, mais également, réduire l'adaptation au canal à une simple soustraction spectrale de la moyenne du signal.

Par ailleurs, l'utilisation d'un GPU pour le calcul du DNN a aussi permis d'obtenir plus rapidement l'ensemble des probabilités des états. Il ne s'agit plus désormais d'un facteur handicapant, voire bloquant, pour obtenir un décodage rapide [2].

2.2.2 Algorithme de décodage

Deux techniques peuvent être utilisées pour le décodage. Tout d'abord, la plus classique consiste en une utilisation séparée des 3 sources de connaissances, associée à la construction du parcours, qui se se fait "à la volée". La seconde, introduite en 1996, utilise le formalisme des FST. En ayant recours à ce formalisme, chaque source de connaissances peut être mise en forme. Les FST permettent de calculer le coût d'une transformation d'une chaîne en une autre. Ainsi, la figure (1) illustre l'estimation du coût par un FST (0,5+2,5+3,5) de la transformation de la séquence *ac* en la séquence *xz*.

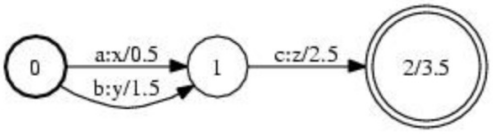

Figure 1 – exemple de FST

On utilise ensuite les algorithmes de composition et d'optimisation de ces mêmes FST pour calculer un unique FST représentant l'ensemble de ces connaissances. Il ne restera donc plus qu'à le parcourir au moment du décodage. En effet, c'est lui qui va permettre de calculer le coût (donc la probabilité) de passer d'une suite de trames acoustiques à une séquence de mots.

Toutefois, le problème majeur de cette démarche reste la taille du FST. Ce dernier devient vite impossible à gérer dans le cas de modèle de langage d'un ordre strictement supérieur à 2 et d'une taille de vocabulaire supérieur à quelques milliers de mots. La figure (2) représente le FST

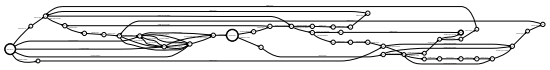

FIGURE 2 – FST complet pour un vocabulaire de deux mots et un modèle unigramme

engendré pour un vocabulaire de 2 mots et un modèle de langage unigramme. On ne peut alors que constater sa complexité. Pour pallier ce problème, le principe est de faire un décodage avec un modèle 2-gramme (la probabilité d'un mot ne dépend que du précédent), d'obtenir ensuite un graphe de solutions puis de le réévaluer par un modèle n-gramme d'ordre supérieur. À titre d'exemple, la taille du FST pour un modèle 2G et un vocabulaire de 160 Kmots est de 11 Go.

Le recours à la boite à outils Kaldi [3], qui utilise à la fois la répresentation en DNN pour l'acoustique et les FST pour les autres sources de connaissances, permet d'obtenir un décodage non seulement performant mais aussi rapide : (1/20) du temps sur une machine équipée d'un GPU et d'un processeur à 20 threads dans un traitement par lot. Avec de tels procédés, la transcription est offerte dans des délais en deçà du temps réel. Toutefois, la difficulté est de s'approcher au plus près d'une telle une rapidité de réponse, c'est-à-dire en évitant les délais trop importants, lorsque le décodage est effectué parallèlement à la prise de son. Pour ce faire, un système dit à latence faible a été mis au point.

2.3 Décodage à latence faible

Dans le cas d'une transcription simultanée, le principal critère à respecter est la latence. Cette notion correspond au temps entre le moment où le locuteur prononce le mot et le temps où ce mot est effectivement affiché. Une première version d'un tel système, fondé sur KALDI, a été proposé par [4]. Dans cette version, l'auteur utilise des modèles GMM mais il n'a pas réalisé un rescoring performant. Cela limite à la fois la taille du vocabulaire et la taille des GMM. De nombreuses modifications s'avèrent donc nécessaires pour arriver à un système à latence faible performant.

2.3.1 Évolution de la Modélisation Acoustique

Pour obtenir un système à latence faible performant, la première modification concerne les GMM : il convient de les remplacer par un DNN qui peut largement être calculé en utilisant la carte graphique d'un ordinateur portable. Ainsi, on peut déléguer les cœurs du CPU aux autres tâches. Le deuxième champ d'adaptation est la gestion de la soustraction spectrale de la moyenne du signal. Il convient d'utiliser une fenêtre glissante pour effectuer cette moyenne. Toutefois, il faut veiller à développer une méthode pour ne calculer cette moyenne que sur les zones où l'on est certain que le locuteur parle. Le risque réside dans la perturbation de la représentation spectrale si l'on intègre à cette moyenne trop de zones sans parole.

2.3.2 Réévaluation des graphes de décodage

La nécessité d'utiliser un vocabulaire de taille suffisante (environ 60 Kmots) impose d'effectuer un premier décodage avec un modèle de langage d'ordre 2. Les raisons sont doubles : d'une part, il faut tenir compte de la taille mémoire. D'autre part, il convient de ne pas négliger le critère rapidité. Ensuite, une réévaluation par un modèle d'ordre supérieur est effectuée. Toutefois, la technique habituelle utilisée dans Kaldi, qui se fonde sur la composition des FSM et leur optimisation, est trop consommatrice de ressources pour effectuer cette tâche en latence faible. En effet, il n'est possible de commencer le travail que lorsque l'on a estimé que le dernier mot de la sentence a été prononcé et décodé. Ce processus doit donc être très rapide. C'est la raison pour laquelle il a fallu développer un décodeur spécifique pour effectuer cette tâche. Ce décodeur parcourt en parallèle les deux FST : le premier représente le graphe de décodage tandis que le second correspond au FST qui prend en considération le modèle de langage d'ordre supérieur pour obtenir la meilleure solution compatible.

2.3.3 Gestion des sorties du décodage

L'ordre des modèles de langage n'est pas sans conséquence sur le processus de décodage. L'on ne peut être sûr qu'un mot est le plus probable que quand on est en train de décoder les tranches de signal correspondant aux n+1 mots suivants, n étant l'ordre du modèle de langage : il vaut 3 pour ce type d'applications. À partir de là, deux choix s'offrent à nous.

Tout d'abord, il est possible, pour arrêter le décodage, d'attendre la détection par le décodeur d'une zone sans parole et/ou éventuellement d'une zone de rupture dans la construction de l'énoncé. On procède alors à sa réévaluation de manière indépendante pour ensuite lancer un nouveau décodage sur les trames qui arrivent. Les critères d'arrêt seront alors pondérés par le temps qui s'est écoulé depuis la dernière fois où on a relancé le décodage. Néanmoins, dans la mesure où la réévaluation par le modèle de langage est effectuée sur ces périodes, il ne faut pas qu'elles soient trop courtes ni que les coupures aient lieu au mauvais moment sous peine de nuire au décodage tout entier.

Pour pallier ces inconvénients, une seconde méthode a été developpée. Un intervalle de temps de l'ordre de 3 secondes est fixé *a priori*. À chacun de ces intervalles, on effectue une réévaluation du graphe qui est en train d'être construit par le décodage du signal. N'est alors affichée que la partie de la solution qu'on estime stable, c'est-à-dire celle qui correspond à plus d'une seconde de la fin de la partie réévaluée. Afin de ne pas saturer la mémoire, il faut aussi de temps en temps réinitialiser complétement le décodage sur le signal. Le temps n'étant plus critique, il suffit alors d'attendre une pause caractéristique dans le discours.

Les deux méthodes présentent des intérêts. Le choix dépend de l'objectif visé dans l'utilisation du système ASR. S'il ne s'agit uniquement que de transcription, la seconde méthode est la plus agréable car les mots sortent avec la latence la plus faible puisque celle-ci est bornée. Si l'on choisit de procéder en plus à une traduction de la parole, la première est préférable car elle permet de confier au système de traduction, au prix d'une latence plus élevée, des syntagmes plus complets facilitant ainsi leur traduction en contexte.

3 L'Adaptation : Étude de cas

Ces problèmes concernant la modélisation acoustique et la latence étant pris en considération, il devient possible d'offrir une transcription qui s'affiche presque immédiatement après le flux de paroles du locuteur. Cependant, outre ces difficultés d'ordre technique, il est également important de considérer les enjeux d'un tel travail.

3.1 Les Enjeux

3.1.1 La Parole Spontanée

Tout d'abord, deux types de parole peuvent être considérés dans cet exercice de transcription. D'une part, l'on peut être confronté à une parole préparée, presque lue, où le locuteur s'est approprié son discours. Il le maîtrise alors suffisamment pour être un orateur efficace dont la parole est fluide, sans heurts. D'autre part et à l'inverse, il peut s'agir d'un discours pris sur le vif, sans préparation aucune (interview en micro-trottoir par exemple). Le locuteur fait alors montre d'une parole spontanée, plus ou moins hâchée, et non exempte de disfluences en tout genre (des *heu* et autres pauses sont disséminées dans tout le propos). Ces deux extrêmes sur l'échelle de la parole illustre tout la difficulté du travail de transcription. Plus la parole sera spontanée, plus les marques d'émotion, de fébrilité seront présentes et plus la construction des énoncés en pâtira.

Le système ASR du LIUM, à faible latence, a été pensé pour permettre la transcription de discours officiels où les discours sont travaillés et où les orateurs sont des habitués de l'exercice. La parole est spontanée mais le degré de spontanéité demeure circonscrit à des niveaux appréhendables. En outre, il s'agit d'un système généraliste prévu pour avoir la couverture la plus large possible en termes de vocabulaire et de structures morphosyntaxiques.

3.1.2 Adaptation et Données d'apprentissage

Pour permettre un tel système, nous l'avons vu, il est nécessaire d'avoir de nombreuses données d'apprentissage pour construire le Modèle de Langage. Ce dernier est donc appris sur un ensemble de corpus écrits, issus soit de journaux écrits comme *Le Monde*, soit de transcriptions manuelles ou automatiques (reçues lors de campagnes d'évaluation par exemple). Il subit également une phase d'optimisation et de pondération pour tenter de prédire au mieux les textes présents dans un corpus de dévelopement qu'on espère proche de ce que l'on aura à reconnaître. Enfin, l'absence de certains mots non rencontrés dans les corpus d'apprentissage pose problème dans la mesure où la bonne reconnaissance de leur suite phonétique ne pourra donner lieu à la production de la suite graphémique du fait même de cette absence. Les mots Hors Vocabulaire (dorénavant OOV) constituent une difficulté majeure pour les système ASR. En conséquence, il est important de garder à l'esprit que la tâche n'est pas triviale, même pour un système robuste et générique comme celui du LIUM.

Par ailleurs, il peut s'avérer nécessaire de procéder à des adaptations pour permettre au système de rendre compte de la réalité d'un discours qui, outre son caractère plus ou moins spontané, fait montre non seulement d'un sujet dit "de spécialité" mais aussi d'un vocabulaire terminologique. Ce fut le cas lorsque le LIUM s'est vu confier la tâche de transcrire une soutenance de thèse en acoustique. L'objectif était de réaliser le sous-titrage de la soutenance afin de permettre aux personnes malentendantes présentes dans la salle de suivre la présentation et les débats qui en découlaient.

3.1.3 Une soutenance de thèse : caractérisation de la parole

Une soutenance de thèse est un exercice spécifique où la prise de parole répond à des codes identifiables : tout d'abord, le futur docteur présente de manière globale son travail puis, dans un second temps, répond aux questions des membres de son jury. Ces deux temps forts font état d'une parole plus ou moins spontanée, qui porte les traces de cette situation d'énonciation particulière. Le candidat, seul orateur dans ce monologue imposé, utilise une langue peu spontanée dans la mesure où l'exposé a été répété. En outre, son discours est soutenu par les diapositives qui lui permettent de conserver les mots-clés et de leur faire la part belle. En revanche, l'émotion de sa voix se fait souvent ressentir et peut entraîner la présence de disfluences et/ou générer des lapsus. Enfin, un vocabulaire terminologique est largement utilisé et les constructions morphosyntaxiques idoines sont souvent répétées.

Lors du passage aux questions du jury, la situation d'énonciation change puisque la parole est partagée entre le candidat et les membres de son jury. Il faut d'ailleurs noter que ce partage de la parole peut s'avérer délicat dans la mesure où les tours de parole ne sont pas toujours clairement identifiés et où les membres du jury entremêlent souvent leur voix. Il devient parfois difficile d'identifier qui parle, quand et à qui. En outre, la présence d'un micro longue distance accroît les difficultés techniques. La reconnaissance de la parole dans ce milieu ambiant particulier, associé à l'utilisation de constructions morphosyntaxiques idoines et d'un vocabulaire terminologique rendent nécessaires une adaptation du système générique pour lui permettre de capter et de taiter au mieux le flux de paroles en entrée.

3.2 La Métrique utilisée : le WER

Pour mesurer la qualité et la robustesse d'un système ASR, on utilise une métrique spécifique, appelée le Word Error Rate (WER). Cette dernière permet d'estimer la capacité d'un système à transcrire au plus près des paroles prononcées en calculant les erreurs commises par ce dernier. Pour être performant, un système doit donc être au plus près de 0 en WER. À l'heure actuelle, un système performant est aux alentours des 10 à 15% de WER pour des discours en parole relativement spontanée, du type de ceux rencontrés sur les chaînes d'actualités où les professionnels de la parole côtoient des invités moins habitués aux plateaux de télévision.

Le WER est une métrique fondée sur la distance de Levenshtein, utilisée dans l'outil Sclite (`http://www.nist.gov/itl/iad/mig/tools.cfm`). Celle-ci est calculée sur les mots. Elle permet, après alignement de l'Hypothèse (la sortie du système) sur la Référence (la transcription manuelle produite par un humain-expert) de calculer les mots corrects mais aussi les substitutions (un mot remplacé par un voisin), les suppressions (un mot en moins dans l'hypothèse) et les insertions (un mot en plus

dans l'hypothèse). C'est en faisant la somme de ces trois erreurs qu'est calculé le pourcentage de WER.

On l'a vu, outre l'adaptation du modèle acoustique et la réduction de la latence, une soutenance de thèse pose des problématiques spécifiques qu'il convient de régler pour rester dans des WER corrects et performants et pour permettre à tous de ne rien manquer de ce moment.

4 Le système issu des adaptations

La méthode générale pour adapter un système est de disposer d'un corpus de développement qui est censé "ressemblé" au signal que l'on devra transcrire. On peut ainsi sélectionner au mieux les mots qui vont constituer le vocabulaire, ajuster les pondérations et autres seuils. Or, pour cette expérience originale, nous ne disposions pas, au préalable, d'un tel corpus. Il a donc fallu effectuer une adaptation à l'aveugle. Une fois l'expérience réalisée, nous avons pu transcrire manuellement une grande partie de la soutenance en nous appuyant sur la transcription automatique, ce qui nous a permis de mesurer a posteriori l'apport des différentes adaptations.

4.1 Données disponibles

Parmi tous les corpus textuels disponibles, nous ne disposons que du seul manuscrit de la thèse d'une part et, d'autre part, d'un corpus provenant des transcriptions en français de TED (`https://www.ted.com/`). Ce dernier a l'avantage de présenter le même type de discours en soliloque de la part d'un expert du domaine, toutefois non rompu à l'exercice oral. En revanche, nous n'avions pas eu accès à des articles du domaine et/ou à d'éventuelles transcriptions de répétitions de soutenance ou autres colloques. Bien entendu, nous avions également à notre disposition les corpus servant habituellement à construire le modèle général audio. Il s'agit des transcriptions d'émissions radiophoniques, d'articles du *Monde* allant de 1988 à 1994 et de quelques GoogleNews. Le tableau 1 indique le nombre de mots présents dans chacun des corpus.

Tableau 1 – Taille des corpus

Corpus	#mots
Thèse	32 217
Transcriptions	4 10^6
Le Monde	316 10^6
GoogleNews	79 10^6

4.2 Vocabulaire, Modèle de Langage et Dictionnaire

4.2.1 Le vocabulaire

La première chose à effectuer est de sélectionner un ensemble de mots qui constituera le vocabulaire. La stratégie consiste à choisir les 60 000 mots les plus fréquents du vocabulaire habituel, auquel on a ajouté tous les mots présents dans le manuscrit de thèse. A posteriori, on peut regarder le pourcentage des mots de la transcription absents du vocabulaire (Tableau 2). Le décompte peut être fait de deux manières différentes : tout d'abord, on peut compter les occurences. Par ailleurs, on peut se référer aux mots distincts. La première méthode prend en compte le nombre de répétitions des mots. Elle est donc corrélée avec le nombre d'erreurs puisque chacune de ces occurrences provoque au moins une erreur.

Tableau 2 – Pourcentage des Mots Hors Vocabulaire (OOV)

Partie	OOV Occurences		OOV mots	
	Stand.	Adap.	Stand.	Adap.
Discussion	1,34%	1,26%	3,71%	3,83%
Soutenance	1,65%	0,61%	4,14%	3,05%
Ensemble	1,53%	0,88%	4,34%	4,20%

Ainsi, on peut remarquer en observant le tableau 2 que le taux d'OOV entre les deux modèles est assez proche. Dans ce tableau (comme dans le tableau 3), le terme *Soutenance* désigne l'exposé de la candidate, tandis que le substantif *Discussion* fait référence aux questions du jury. Le vocable *Ensemble* désigne bien évidemment la réunion des deux.

On peut tout de même constater que les différentes adapations ont eu un impact sur la transcription de l'exposé de la candidate, puisqu'on passe de 1,65% à O,61% pour les occurences et de 4,14% à 3,05% pour les mots. Une analyse a posteriori a permis de constater une meilleure représentation des mots propres au domaine par le modèle adapté. Ceci n'est pas très perceptible quantitativement en termes de WER mais se perçoit aisément qualitativement lorsqu'on lit la transcription.

4.2.2 Construction du Modèle de Langage

Le principe de la construction du Modèle de Langage est, une fois le choix de vocabulaire effectué, d'en construire un pour chacun des corpus. Le modèle final est issu d'une interpolation de ces modèles. Comme indiqué précédemment, les poids sont calculés par optimisation de la perplexité sur un corpus de développement. La perplexité correspond à l'inverse de la moyenne géométrique des probabilités des mots du corpus estimées par le modèle. Plus la perplexité est faible, meilleure est l'adéquation du modèle de langage au discours à transcrire. Nous avons donc décidé de donner a priori un fort poids aux deux corpus que nous avons estimé les plus proches de ce devrait être la soutenance : le texte de la thèse et les transcriptions en français de TED. On peut finalement constater que le démarche adoptée est concluante pour la perplexité, comme le montre l'ensemble des résultats du Tableau 3.

Tableau 3 – Perplexité

Partie	LM orig.	LM adapté
Soutenance	227.48	112.49
Discussion	179.76	130.25
Ensemble	232.22	119.42

4.2.3 Le dictionnaire

Le dictionnaire constitue l'une des 3 bases de connaissances qu'utilise un système ASR. Il a pour fonction de faire correspondre le Modèle Acoustique et le Modèle de Langage en proposant pour chaque mot du vocabulaire une ou plusieurs suite de phonémes caractérisant sa prononciation. Toutefois, les outils de phonétisation utilisés pour permettre "la traduction" d'un graphème en une suite de phonèmes respectant sa phonétique mais aussi sa phonologie reçoivent en entrée des mots normalisés, puisque issus du Modèle de Langage, c'est-à-dire sans majuscule et respectant l'encodage Latin1. Il n'est donc pas rare que cette normalisation engendre des erreurs. En effet, les outils se fondent également sur la casse pour établir qu'ils ont affaire à un nom propre. Ils tentent aussi par ce biais de déterminer la langue d'origine de la lexie. Toutefois, une analyse fine des sorties de transcriptions ne fait pas apparaître d'erreurs uniquement dues à la phonétisation ratée de mots. Ainsi, le nom propre Hammerstein, largement cité dans la présentation de la thèse (au sein de l'expression "modèle de Hammerstein") est suffisamment discriminant phonétiquement pour être correctement phonétisé par les outils automatiques. On peut conclure a posteriori, qu'aucun retravail manuel (vérification et/ou corrections) sur les phonétisations ne s'imposait.

5 Résultats et Discussion

Il convient à présent d'examiner les résultats des expériences. Ces résultats ont pu être calculés a posteriori puisque les premières adaptations (choix des données et pondération des Modèles de Langage) ont été faites à l'aveugle. Il s'agit de comparer les résultats en WER du système générique par rapport au système ayant permis la trancription de la thèse. Puis, dans un second temps, nous comparerons les résultats du système dédié avec les résultats issus de nouvelles adaptations, réalisées bien après la soutenance puisque nous possédions une transcription de référence, réalisée par un transcripteur humain expert. Au sein des tableaux de résultats, la ligne intitulée *Base* fait référence au système générique tandis que la ligne appelée *Balb* renvoie au système dédié. En outre, le triptyque Soutenance/Discussion/Ensemble (abrégé en Sout/Disc/Ens) désigne la même partition de la soutenance que celle utilisée pour les tableaux précedemment cités.

5.1 Système générique *vs* système dédié

5.1.1 Résultats

Tableau 4 – Adaptation des systèmes

Exp.	#mots	Cor.	Sub.	Sup.	Ins.	WER
Base.Sout	5563	81,9	14,0	4,0	2,7	**20,8**
Balb.Sout	5563	87,0	9,7	3,2	1,8	**14,7**
Base.Disc	3800	59,8	17,3	22,6	2,1	42,0
Balb.Disc	3800	62,1	16,4	21,3	2,0	39,7
Base.Ens	9363	72,9	15,4	11,6	2,5	**29,4**
Balb.Ens	9363	76,9	12,4	10,6	1,9	**24,9**

On s'aperçoit que le modèle adapté, originellement architecturé pour tenter de correspondre au mieux au discours supposé de la soutenance, et notamment de la présentation répond effectivement aux attentes puisque les écarts sont importants. En effet, on diminue le WER de 6,1 pt pour la présentation des travaux de thèse, et de 4,5 pt pour l'ensemble de la soutenance. On peut notamment constater que, parmi les erreurs permettant de mesurer le WER, les substitutions diminuent largement lors de la présentation (passant 14,0 à 9,7). En revanche, les scores restent élevés lors de la partie Discussion avec le jury. Ce phénomène s'explique par l'utilisation quelque peu anarchique du micro et par une répartition désordonnées de la prise de parole, les membres du jury discutant entre eux tout autant qu'avec la candidate.

5.1.2 Discussions

Il apparaît donc que le système adapté qui comporte un Modèle de Langage dédié, construit par pondération sur le manuscrit de la thèse et des discours de même type (discours scientifiques en soliloque), est efficace. Il permet de réduire le WER et d'offrir une transcription très correcte, notamment pour la partie présentation des travaux. Toutefois, l'examen minutieux mené a postériori fait état de quelques erreurs résiduelles auxquelles il est possible d'apporter des solutions rapides.

5.1.3 Quelques pistes d'exploration

Les corrections à apporter concernent deux principaux champs d'action. D'une part, il s'est avéré que la normalisation d'un vocable portant un trait d'union était à revoir. En effet, le mot *haut-parleur*, bien que très présent dans la présentation (on compte 38 occurrences au total), n'était pas toujours bien reconnu et faisait l'objet de substitutions. D'autre part, une phase d'adjudication, qui vise à corriger les erreurs faites dans la référence, offre la possibilité de ne pas pénaliser le système lorsqu'il n'est pas à l'origine de l'erreur.

5.2 Le système dédié : réévaluation après corrections

Après application des corrections proposées, à savoir une normalisation constante du vocable *haut-parleur* et une relecture attentive de la référence, nous avons procédé à une réévaluation de la transcription.

5.2.1 Résultats

Dans le Tableau ci-après les préfixes *HP.* et *HP+Adj* renvoient respectivement à la correction portant sur le Haut-parleur puis à l'ensemble des corrections apportées, incluant la phase d'adjudication. Les résultats bruts du système dédié sont rappelés grâce au préfixe *Balb*.

Le simple fait de modifier la normalisation du mot *haut-parleur* en corrigeant l'apprentissage de la synapsie permet de légèrement faire descendre le WER. Mais c'est surtout la phase d'adjudication qui permet de rétablir la justesse du WER, en n'imputant pas au système des erreurs qui ne sont pas de son fait.

Tableau 5 – Corrections : Normalisation et phase d'Adjudication

Exp.	#mots	Cor.	Sub.	Sup.	Ins.	WER
Balb.Sout	5563	87,0	9,7	3,2	1,8	14,7
HP.Sout	5563	87,3	9,6	3,1	1,8	14,5
HP+Adj.Sout	5563	88,6	8,4	3,0	1,7	**13,1**
Balb.Disc	3800	62,1	16,4	21,3	2,0	39,7
HP.Disc	3800	62,6	16,5	20,7	2,1	39,3
HP+Adj.Disc	3784	63,6	16,0	20,2	2,0	**38,2**
Balb.Ens	9363	76,9	12,4	10,6	1,9	24,9
HP.Ens	9363	77,3	12,4	10,2	1,9	24,5
HP+Adj.Ens	9347	78,4	11,5	10,0	1,8	**23,3**

5.2.2 Discussion

Outre la normalisation du vocable *haut-parleur*, c'est l'adjudication qui permet de descendre le WER sur les sorties de transcriptions puisque l'on passe, par exemple, de 14,7% à 13,1% sur la présentation des travaux par la doctorante.

La phase d'adjudication correspond à une étape importante qui permet de corriger, non pas les sorties de système, mais la référence elle-même. En effet, il subsiste toujours des erreurs dans un travail manuel et il peut arriver que le système soit mis en défaut alors que c'est précisément lui qui donne la bonne solution. Puisque ce dernier fonctionne par apprentissage, il restitue ce qu'il a appris dans les corpus. Ainsi, si le transcripteur commet des fautes d'orthographe, le système, lui, ne les reproduira sans doute pas et se retrouvera injustement pénalisé.

Hormis les fautes d'orthographe, un autre type d'erreurs survient : il concerne certains accords. En effet, il est parfois autorisé en français de procéder ou non à l'accord en fonction des liens logiques que l'on souhaite mettre en relief. Par exemple, dans la soutenance de thèse, la doctorante évoque un *guide d'onde* que la transcriptrice a écrit *guide d'ondes* avec -s tandis que le manuscrit de la thèse fait état d'un *guide d'onde*. Le système a donc restitué la version apprise sans -s. Il semble alors inapproprié de compter une substitution.

Enfin, il faut souligner que la spécificité d'une telle transcription requiert une culture scientifique certaine et qu'il est parfois difficile de transcrire la réalité de ce qui est dit. C'est notamment le cas pour les descriptions et explications d'algorithmes et pour les descriptions d'expérimentation, le matériel utilisé étant souvent inconnu des béotiens du domaine.

5.2.3 De nouvelles pistes d'amélioration

Outre ces corrections a posteriori qui permettent au modèle dédié de révéler ses véritables possibilités, il apparaît que des erreurs d'accords subsistent dans les sorties de transcriptions. Par exemple, le syntagme *les éventuelles résonnances* fait l'objet d'une faute d'accord et est transcrit *les éventuelles résonnance*. L'explication est simple : ce trigramme n'a jamais été appris par le système car il est absent des corpus. Dans ce cas précis, le modèle de langage effectue un repli et la probabilité du syntagme est fonction de la probabilité du dernier mot. Or *résonnance* est plus courant dans les corpus que *résonnances* et c'est donc *résonnance* qui est choisi. Ceci est plus fréquent dans ce cadre du fait de la faible disponibilité de corpus de spécialité. Pour pallier ces phénomènes d'accords tronqués liés à des modèles à classe qui ne fonctionnent pas, une expérimentation à partir de Réseaux de Neurones Profonds est en cours. Toutefois, les résultats déjà obtenus, du moins pour la partie Discussion, démontrent qu'il est non seulement possible d'obtenir des systèmes ASR robustes mais aussi dédiés à la tâche.

6 Conclusion

Au terme de cette analyse, on constate que les systèmes ASR, d'habitude utilisés dans le cadre d'une évaluation stricte lors de campagnes nationales ou internationales, peuvent également faire l'objet d'adaptation afin de correspondre à la réalité des propos tenus. Lors de campagne d'évaluation, c'est la robustesse qui est avant tout recherchée. Les systèmes présentés demeurent sont alors optimisés sur un corpus de développement, dont on espère qu'il sera proche des données d'évaluation. En revanche, lorsqu'un cadre applicatif spécifique et terminologique est donné, il est possible de procéder à des adaptations. Ces dernières peuvent tout aussi bien concerner les conditions de décodage que les types de données à traiter : l'acquisition de données idoines, leur pondération, la prise en compte du contexte ou bien encore la taille du vocabulaire sont alors recherchées. On constate alors une amélioration constante des résultats au fur et à mesure des améliorations apportées au système.

Remerciements

Nous tenons ici à remercier Madame Balbine Maillou, docteur en acoustique, qui nous a contacté pour sous-titrer sa soutenance de thèse, intitulée "Identification et caractérisation des non-linéarités de systèmes acoustiques" et qui nous a permis de mener à bien ce travail.

Références

[1] D. Erhan, Y. Bengio, A. Courville, P. A. Manzagol, P. Vincent and S. Bengio, Why Does Unsupervised Pre-training Help Deep Learning ?, *Journal of Machine Learning Research*, **11**, (2010).

[2] H. Hermansky, S. Sharma, Temporal patterns (TRAPS) in ASR of noisy speech, *Acoustics, Speech, and Signal Processing, IEEE International Conference*, **1**, 289-292 (1999).

[3] K. Vesel, A. Arnab, L. Burget, D. Povey, Sequence-discriminative training of deep neural networks, *INTERSPEECH*, 2345-2349, (2013).

[4] T. Alum, Full-duplex Speech-to-text System for Estonian, *Baltic HLT 2014*, (2014).