Résumé long

Introduction

Les modèles de prononciation, comme plus gnralement la reconnaissance de la parole (automatique speech recognition (ASR)), ont connu une évolution progressive des types de données traitées, du discours lu au discours spontané, accompagne d'une évolution des méthodes utilisées depuis les méthodes fondées sur des connaissances linguistiques aux méthodes principalement statistiques. Malgré une longue histoire de la recherche dans ce domaine, il existe toujours un vif intérêt pour améliorer la construction du lexique de reconnaissance surtout pour des applications qui impliquent le traitement de parole spontanée ou moins canonique (par ex., accentuée).

La variation de prononciation peut être modélisée aux différents niveaux d'un système de reconnaissance de la parole, comme expliqué dans (Strik and Cucchiarini, 1999): au niveau du lexique, du modèle acoustique, du modèle de langage (LM). Au niveau acoustique, des phones dépendants au context représentent les variations dans des contextes particuliers. Au niveau du lexique, des prononciations multiples par mot peuvent être incluses pour représenter explicitement la variation de prononciation. Au niveau du modèle de langage (LM), les variations inter-mots des prononciations sont traitées avec des réseaux de grammaire, des LM statistiques ou des modèles multi-mots.

Dans cette thése, on étudie la représentation de la variation phonémique au niveau du lexique de façon explicite, en ajoutant les variantes de prononciation. La premiére partie de ce travail est consacrée à la génération automatique des prononciations et des variantes de prononciation pour le lexique d'un système de reconnaissance de la parole. Une application directe de ce travail consiste à prédire la prononciation des mots hors-vocabulaire (out-of-vocabulary words (00V)). Une autre application consiste à enrichir le dictionaire avec des variantes de prononciation de haute qualité, qui peuvent améliorer la performance du système si elles sont traîtées proprement (par ex. entraînement de leurs poids, pruning,...). Ces variantes sont souvent nécessaires, surtout lorsque l'on travaille avec de la parole conversationnelle où des formes de prononciation réduites remplacent souvent les formes canoniques. Ensuite, on se focalise sur les phenomènes de confusabilité introduits dans le système lorsque l'on augmente le lexique de prononciation. Des efforts seront réalisés dans le sens de la recherche d'une mesure qui représente bien cette confusabilité et possède idéalement une bonne corrélation avec le taux d'erreur du système. Une méthode d'entraînement disriminant est proposé pour ajouter des variation phonémiques au lexique en contrôlant les effets négatifs de la confusabilité. Ces methodes sont etudiées sur deux tâches: la reconnaissance automatique de la parole et la détection des mots-clés (keyword spotting (KWS)). Des fonctions objectifs différentes sont choisies pour chaque tâche, afin d'optimiser chaque fois la mesure de performance correspondante.

La conversion graphème-phonème (g2p) est la tâche consistant à trouver la prononciation d'un mot partir de sa forme écrite. Malgré plusieurs décennies de recherche, cette tâche reste difficile et joue un rôle dans de nombreuses applications des technologies du

langage humain. C'est un élément essentiel des systèmes de reconnaissance et de synthèse de la parole. En outre, avec l'utilisation à grande échelle des données réelles, il existe des mots ne figurant pas dans le dictionnaire de reconnaissance (mots hors-vocabulaire), pour lesquels une prononciation généré rapidement et automatiquement est souvent nécessaire. Ces mots sont souvent des entités nommées qui dépendent fortement de l'actualité et, par conséquent, il est difficile de prédire leur présence dans les discours et de les inclure dans le dictionnaire (Béchet and Yvon, 2000). Un autre domaine d'application de la tâche de phonétisation dans le traitement de langues est la détection et la correction des erreurs orthographiques (van Berkel and De Smedt, 1988). De plus, la relation étroite entre la phonologie et la morphologie est bien connue et des phénomènes morphologiques d'origine purement phonologique ou guidés par des contraintes phonologiques sont bien étudiés (Kaisse, 2005). D'autres applications comprennent l'apprentissage des langues assisté par ordinateur, l'apprentissage des prononciations et les systèmes d'"e-learning" en général.

La conversion graphème-phonème est un moyen de générer des prononciations totalement indépendamment d'un ensemblede données de parole. Le modèle de prononciation est souvent la seule partie d'un système de reconnaissance de la parole qui est statique et qui n'est pas formé d'un ensemble de données particulier. Cependant, le besoin d'un dictionnaire dynamique adaptée aux données de parole est jugé important pour améliorer les performances de la reconnaissance de la parole. L'ajout d'un grand nombre de variantes à un dictionnaire de reconnaissance, même si ces variantes sont de très bonne qualité, ne conduit pas nécessairement à une amélioration du taux d'erreur du système. Les prononciations alternatives introduisent dans le système de la confusabilité qui est difficile à mesurer et de limiter d'une manière appropriée. Ajouter des prononciations alternatives avec des probabilités peut modérer cette confusabilité, mais l'apprentissage de ces poids de prononciation reste un problème ouvert.

La confusabilité introduite par des multiples prononciations par mot est liée à l'augmentation du taux d'homophones (des mots qui se prononcent de la même manire mais qui sont écrits différemment), ce qui signifie que ces variantes supplémentaires peuvent ne pas être utile à la performance de la reconnaissance (Tsai et al., 2001). Un exemple typique en anglais est le mot "you": la prononciation canonique /yu/ est choisie lorsque une seule variante est utilisée; la modélisation de variation nécessite de considérer les prononciations /yu/ et /yc/, qui sont toutes les deux présentes dans notre dictionnaire. Cette dernire prononciation (/yc/), dans la phrase "you are", est facilement confondu avec /ycr/, la prononciation du mot "your". Ces confusions, en particulier quand elles impliquent des mots fréquents, peuvent causer une dégradation du système de reconnaissance quand plus d'alternatives sont ajoutés.

Pour gérer cette confusabilité, il a été montré utile de faire usage également de l'information acoustique des données de parole, afin de choisir les variantes à ajouter dans le lexique de reconnaissance. De cette façon, un lexique dynamique, adapté aux données peut être construit. Ce lexique est idéalement entraîné pour minimiser le taux d'erreur du système, comme cela est fait traditionnellement pour l'entraînement du modèle acoustique et est aussi essayé pour l'entraînement du modèle de langage.

Génération automatique des prononciations

Dans une première partie, deux méthodes sont proposées, inspirées par la traduction automatique, afin de dériver des prononciations et des variantes de prononciation d'un lexique initial. Tout d'abord, un outil de traduction automatique, Moses (Koehn et al., 2007), est utilisé comme un convertisseur g2p pour extraire une liste de n-meilleures prononciations. Le même outil est utilisé pour effectuer la conversion phonème-phonème (p2p) et d'en tirer des variantes d'une prononciation connue canonique. La deuxième approche est une nouvelle méthode basée sur une approche "pivot", utilisée traditionnelement pour la tâche d'extraction de paraphrases et appliqué ici dans une étape de post-traitement au convertisseur g2p ou directement à la prononciation canonique comme un convertisseur p2p. Notre convertisseur g2p permet de générer des prononciations pour des mots hors-vocabulaire, mais également d'enrichir le dictionnaire d'origine avec des variantes de prononciation, tandis que les convertisseurs p2p se concentrent sur l'ajout de variantes au dictionnaire.

Les performances de ces deux méthodes sont testées pour l'anglais sous différentes conditions d'entraînement (en utilisant un lexique avec ou sans variantes de prononciation). Les résultats de rappel et de précision pour les tâches de conversion g2p et p2p prouvent la génération de prononciations de bonne qualité. Une observation générale est que les n-meilleures listes de Moses ont un rappel plus élevé quand la comparaison est faite pour toutes les prononciations de référence dans les données de test. Cependant, la méthode basée sur pivot génère davantage de variantes correctes. C'est un avantage de la méthode pivot qui pourrait être utile dans certains cas, par exemple pour générer des variantes à partir de la sortie d'un système g2p à base de règles qui, initialement développé pour la synthèse de la parole, ne modéliserait pas les variantes de prononciation, ou pour enrichir un dictionnaire avec des variantes de prononciation limitées.

Lorsqu'elles sont utilisées pour la conversion p2p, ces approches diffèrent dans la maniére dont l'information sur la variation de la prononciation est obtenue. L'approche utilisant Moses pour p2p ne prend en compte que les informations fournies par les transcriptions phonémiques. La méthode pivot utilise des informations provenant à la fois de la transcription phonémique et de la transcription orthographique. Quand le dictionnaire complet (qui contient des mots avec une ou plusieures prononciations) est utilisé pour l'entraînement, la méthode basée sur Moses donne de meilleurs résultats comparativement à la méthode pivot. Les résultats sont similaires quand l'entraînement est effectuée uniquement sur les mots du dictionnaire avec des prononciations multiples. Toutefois, lorsque le dictionnaire d'apprentissage ne contient pas de variantes de prononciation, la mèthode basée sur Moses ne peut plus être utilisée, tandis que la mthode pivot peut encore apprendre à générer des variantes. C'est un avantage de cette méthode qui pourrait être utile pour des langues pour lesquelles il n'existe pas de dictionnaire contenant des prononciations multiples. Ceci découle de l'utilisation de l'information fournie aussi par la transcription orthographique par la méthode pivot.

Les prononciations génér'ees par Moses ont également été utilisées pour effectuer des tests dans un système état de l'art de reconnaissance de la parole. Ces expériences montrent que les prononciations ajoutées sont de bonne qualité, même si un entraînement est

effectué sous conditions de faible variation et qu'elles peuvent améliorer la performance par rapport à un dictionnaire de base sans variantes.

Ces méthodes sont ici testées pour un dictionnaire anglais, mais peuvent être appliqués à d'autres langues puisqu'elles sont totalement automatiques et n'utilisent pas de savoir linguistique. Il faut aussi noter que l'anglais est une langue difficile pour la tâche de conversion graphme-phonème, parce qu'il n'existe pas toujours une correspondance claire entre la forme orthographique et la prononciation d'un mot. Par conséquent, il est difficile de représenter la variation phonémique de la langue en utilisant des règles linguistiques.

Confusabilité phonémique

Une observation concernant le travail sur la géneration automatique des prononciations et en même temps un problème connu de la modélisation des prononciations, est la confusabilité introduite dans le système quand on ajoute des variantes au lexique de reconnaissance. Cette confusabilité, comme déjà mentionné, peut gravement dégrader la performance de reconnaissance de la parole, surtout si les poids des variantes de prononciation ne sont pas convenablement entraînés. Il s'agit alors d'un montant supplémentaire de confusabilité qui est ajouté à la confusabilité "de base" du dictionnaire de prononciation. La confusabilité "de base" est un compromis entre la taille du dictionnaire et le taux des OOVs, ce qui signifie qu'un dictionnaire plus petit en taille (avec moins des mots) aura moins d'homophones et donc moins de confusabilité, mais qu'il présentera un taux supérieur d'OOVs. Par conséquent, il peut présenter une performance pire, cette fois pas à cause du problèmes de confusabilité, mais à cause des mots manquants dans le dictionnaire. Les questions relatives à la mesure et à la réduction de la confusabilité "de base" aussi bien que de la confusabilité supplémentaire lorsque des variantes sont ajoutées sont des problèmes ouverts, parce que les homophones sont un phénomène inhérent de la parole et ne peuvent être ignorées sans nuire à la performance de la reconnaissance de la parole.

Dans cette partie de la thèse, nous nous concentrons sur une compréhension plus profonde de la confusabilité inhérente au dictionnaire de prononciation. En particulier, nous définissons une mesure visant à évaluer à quel degré un modèle de prononciation fonctionnera bien lorsqu'il est utilisé en tant que composante d'un système de reconnaissance automatique. Cette mesure, l'"entropie de prononciation", fusionne les informations à la fois du modèle de prononciation et du modèle de langage et mesure l'incertitude introduite dans le système par ces différentes composantes. Ce score est calculé en composant de façon efficace la sortie d'un reconnaisseur de phonèmes, d'un dictionnaire de prononciation et d'un modèle de langage. Pour ce faire, une représentation fondée sur les transducteurs finis (FST) est adoptée et la bilbiothèque OpenFst (Allauzen et al., 2007) est utilisée. Nous éxperimentons avec cette mesure et différents dictionnaires avec ou sans variantes de prononciation et pondération des prononciations.

On étudie enfin le rôle de cette mesure de confusabilité en tant que facteur prédictif des performances des modèles de prononciation. Cependant, il n'est pas simple et direct de trouver une corrélation entre cette mesure et la performance d'un système de reconnaissance de la parole. Ce qui rend cette procèdure particulièrement complexe, c'est le

fait que les mots confondus sont un phénomène non négligeable de la parole naturelle et le fait de les simplement ignorer réduit significativement la force exhaustive du dictionnaire. Cela signifie qu'un ensemble cohérent de prononciations n'est pas nécessairement relié à un réseau de prononciations de faible perplexité.

L'objectif final de ce travail serait d'essayer d'utiliser cette mesure pendant la procédure de construction du dictionnaire, de manière à enrichir le dictionnaire avec de nouvelles variantes sans dégrader le taux d'erreur du système. Par exemple, on pourrait envisager de trouver un moyen efficace d'intégrer l'"entropie de prononciation" à un objectif qui entraîne convenablement le dictionnaire de prononciation. Ceci fait partie des perspectives de ce travail.

Modèle de confusions phonémiques pour un système de reconnaissance de la parole

Dans la section précédente, nous avons analysé et mesuré la confusabilité de prononciation. Nous avons en particulier observé que cette confusabilité n'est pas directement corrélée avec la performance du système de reconnaissance vocale. Cependant, il est important d'être capable d'ajouter des variantes de prononciation qui peuvent améliorer le taux d'erreur du système, tout en gardant la confusabilité basse. Ayant comme objectif de donner une haute discriminabilité au lexique de prononciation, nous avons choisi d'tudier un cadre discriminant pour l'entraînement des poids des confusions phonémiques de notre système ASR. La fonction objectif à minimiser été choisie pour être le taux d'erreur de phonèmes, qui est directement liée à la performance du système. Dans cette section, nous décrivons le travail sur l'utilisation d'un tel modèle de confusion phonémique, qui élargit l'espace de recherche phonémique par des variantes de prononciation au cours de décodage ASR.

Pour ce faire, nous adaptons un cadre discriminant pour entraîner les poids des modèles de prononciation et nous menons des expériences sur de grand corpus pour évaluer la méthode proposé sur une tâche réelle. Tout d'abord, la sortie d'un reconnaisseur de phonèmes est alignée avec la référence et un ensemble de paires de confusions phonèmiques sont extraites. Ces paires de confusions forment un modèle de confusion qui sera utilisé pour étendre l'espace de recherche phonémique lors du décodage ASR. Pour entraîner les poids de ce modèle de confusion, on le compose avec les lattices produits par le reconnaisseur de phonèmes et on effectue un apprentissage discriminant en minimisant la distance d'édition phonèmique entre les séquences de sortie du reconnaisseur de phonèmes et les véritables sequences phonèmiques de référence. Il en résulte un ensemble de règles phonologiques, qui peuvent être appliquées sur les prononciations de base du lexique. De cette façon, nous espérons avoir des prononciations qui reflètent mieux les séquences réellement prononcées. Il faut noter que, en procédant de cette façon, on n'ajoute pas un nombre fixe de prononciations par mot, contrairement à ce qui est fait avec la conversion g2p statique.

Comme mentionné plus haut, l'apprentissage discriminant est utilisé pour entraîner les poids du modèle de confusions et vise à améliorer directement la performance du système

tout en gardant la confusabilité basse. Dans un modèle discriminant, les paramétres du modèle (c'est-à-dire les probabilités du modèle de prononciation ou de langage) sont adaptées pour minimiser le taux d'erreur de reconnaissance. En revanche, les paramètres d'un modèle du maximum de vraisemblance sont dérivés, comme son nom le suggère, afin de maximiser la probabilité de certaines données connaissant le modèle; une augmentation de la vraisemblance des données d'entraînement ne se traduit donc pas toujours par des taux d'erreur réduits. Une autre façon de voir l'application de notre modèle de confusion est aussi comme un correcteur des erreurs du reconnaisseur de phonèmes. L'étude de (Greenberg et al., 2000) a montré que les erreurs au niveau des phonèmes et au niveau des mots sont corrélées, ce qui justifie notre choix d'une fonction objectif au niveau de phonèmes. Cela nous permet d'ajouter des variantes aux prononciations de base de tous les mots et ne pas se limiter aux mots qui sont présents dans le corpus d'entraînement. En outre, il y a moins de paramètres à entraîner (nombre limité de paires de phonèmes) en comparaison de la recherche des confusions au niveau de mot ou de phrase. Par conséquent, nous sommes capable de travailler avec un système ASR de parole continue segmenté en des longues phrases (env. 80 mots/phrase).

Pour cette étude, on se limite à l'utilisation d'un modèle de confusion phonémique unigramme. Ce travail est présenté comme une preuve que cette méthode et ses extensions possibles (par ex. une généralisation au cas d'un modèle de confusion contextuel est possible) peut être prometteur pour l'adaptation du dictionnaire de reconnaissance sur un ensemble particulier de données. Nous effectuons des expériences en utilisant le perceptron moyenné, initialement proposé par (Freund and Schapire, 1999), et le modèle CRF, initialement proposé par (Lafferty et al., 2001). Les scores acoustiques sont utilisés lors de l'entraînement des poids de prononciation, ce qui nous permet d'intégrer des informations phonémiques fournies par le modèle acoustique. Cela peut améliorer les résultats, comme observé dans (Weintraub et al., 1996) et dans (McGraw et al., 2013). Toute la mise en place est faite avec des transducteurs aux états finis (FST).

Concernant les résultats de ce travail, l'utilisation d'un modl̀e de confusion unigramme n'introduit pas de confusabilité supplémentaire au système quand il est utilisé avec des poids entraînés de façon discriminante. Ceci est une indication que les poids sont appris dans la bonne direction. On peut s'attendre à de nouvelles améliorations lorsque plus de contexte est ajouté au mod´le. Un autre résultat intéressant à souligner est l'amélioration par rapport aux performances de base d'un décodeur FST, tant au niveau des phonèmes qu'au niveau des mots, quand notre modèle de confusions est utilisé lors d'une étape de post-traitement. Cependant, il n'y a pas une solution directe pour intégrer ce modèle de confusion à un décodeur non-FST.

Il convient également de noter que toutes les expériences ont été menées pour des données représentant de la parole continue au sein d'un système ASR grand vocabulaire utilisant une segmentation en longues phrases, ce qui constitue un système de base complexe et difficile à améliorer. De plus, l'anglais est une langue connue pour la difficulté qu'elle présente lorsqu'il s'agit de modélisation au niveau phonémique. À notre connaissance, c'est la première fois qu'une telle méthode discriminante pour apprendre des poids de prononciation est appliquée à un système ASR avec des phrases aussi longues, car une telle tâche présente des difficultés de calcul, ainsi que des difficultés pour analyser les

erreurs ou à observer des améliorations.

Les méthodes proposées peuvent être aussi facilement utilisées au niveau des mots en gardant les mêmes features phonémiques, mais en faisant un décodage au niveau des mots à la place d'un décodage phonémique. Cela permettrait d'ajouter des descripteurs plus complexes, par exemple des traits spécifiques au niveau des mots concernant des transformations phonémiques fréquentes pour certains mots. De cette façon, un contrôle plus global des prononciations générées pourrait être effectué. En outre, la correction des erreurs courantes dans certains mots qui apparaissent fréquemment peut améliorer de beaucoup la performance.

Modèle de confusions phonémiques pour la détection des mots-clés

Dans la section précédente, un modèle de confusion de phonèmes entraîné de façon discriminante a été utilisé pour étendre le lexique de prononciation dans le décodage ASR. Il a été montré que l'apprentissage discriminant des poids des prononciations peut aider à enrichir le lexique sans ajouter de confusabilité indésirable au système. Ensuite, nous avons proposé d'utiliser l'apprentissage discriminant pour la construction d'un modèle de confusion de phonèmes qui élargit l'index phonémique d'un système de détection de mots-clés (KWS) en ajoutant des variations phonémiques.

Les dernières années ont vu croître l'intérêt pour la tâche KWS, une application qui est souvent appliquée comme une étape de post-traitement de la reconnaissance. Comme la quantité de données audio augmente rapidement, la capacité d'y rechercher efficacement certains mots ou expressions devient de plus en plus importante. KWS a pour but de rechercher dans les données audio et d'y détecter un mot-clé donné, qui est généralement un mot isolé ou une courte phrase. Les systèmes KWS fonctionnent généralement en deux phases: l'indexation et la recherche. Le système traite les données audio une seule fois, au cours de la phase d'indexation, sans connaissance des termes de la requête. Cette phase se fait hors ligne et prend le plus du temps. L'index généré est stocké et accessible lors de la phase de recherche, afin de localiser les termes et les lier au son original.

L'indexation au niveau des mots peut sembler être une solution simple, mais il ne peut pas gérer les OOVs qui sont souvent des entités nommées absentes dans le dictionnaire du système ASR. Le taux d'OOVs peut augmenter avec le temps, puisque les dictionnaires sont généralement fixes tandis que les modifications apportées aux données du monde réel se font de façon dynamique. La génération de prononciations pour les OOVs implique d'avoir un système g2p, ce qui n'est souvent pas évident, spécialement pour les langues ayant des ressources limitées. Augmenter le système de reconnaissance de la parole avec des variantes de prononciation peut aider, mais implique la régénération de l'index et peut introduire de confusabilité au système KWS et augmenter les fausses alarmes, surtout si les poids de variantes ne sont pas correctement entraînés. En outre, il existe de nombreuses applications où la performance de la reconnaissance au niveau de mots est fortement dégradée en raison des conditions audio difficiles, et même les mots du vocabulaire sont pas représentés avec succès dans l'index. Pour ces raisons, des systèmes

KWS à base de phonèmes ont été utilisés dans le passé (Saraclar and Sproat, 2004), qui n'imposent pas de restrictions de vocabulaire. Dans le travail actuel, un index phonémique est créé à partir des lattices générés par un reconnaisseur des phonèmes, ce qui nous permet de conserver des informations sur l'incertitude phonémique de la reconnaissance de phonèmes dans l'index.

Cependant, la qualité du reconnaisseur des phonèmes peut être très faible, surtout lorsqu'il s'agit de données enregistrées sous conditions de bruit. Pour cette raison, un modèle de confusions phonémiques est introduit dans ce travail. Une fois appliqué sur l'index, ce modèle de confusion agit comme un correcteur des erreurs de reconnaissance. Notre modèle de confusion augmente l'index avec des séquences de phonèmes alternatifs. Ce fait introduit intrinsèquement des détections supplémentaires lors de la recherche des requêtes. Cela peut être bénéfique, surtout si les contraintes d'espace nous obligent à élaguer de manière significative les lattices de phonèmes afin de maintenir l'index à une taille raisonnable. L'optimisation de KWS doit être une procédure qui augmente les vraies détections tout en gardant le taux des fausses alarmes faible. Pour résoudre ce problme, nous choisissons une méthode d'apprentissage discriminant pour entraîner les paramètres du modèle de confusion, pour laquelle la fonction objectif qui est optimisée est le "Figure of Merit" (FOM), qui est directement liée à la performance KWS. Le FOM a également été utilisé dans (Wallace et al., 2011) afin d'optimiser directement les poids de l'index, qui avait la forme d'une matrice de scores acoustiques probabilistes. Dans (Wang et al., 2009), il a été utilisé pour optimiser un facteur d'interpolation lorsque des variantes de prononciation sont ajoutés pour les OOV. À notre connaissance, c'est la première fois que cet objectif est utilisé pour entraîner les poids d'un modèle de confusion de phonèmes pour la tâche KWS.

Cette approche a été testée pour l'anglais. Cependant, elle est indépendante de la langue et pourrait être appliquée à d'autres langues, y compris pour des langues disposant de ressources limitées, où le problème des OOVs est plus important. Les expériences menées montrent une certaine amélioration de la FOM. Le modèle de confusion est appliqué sur l'index construit en utilisant les lattices de sortie d'un reconnaisseur phonémique très simple (*phone-loop*). Dans le futur, nous envisageons de l'appliquer également des systèmes hybrides mots-phonèmes. De plus, le modèle de confusion utilisée dans ce travail est une modèle unigramme qui ne tient pas compte du contexte phonémique. Notre objectif est d'essayer d'utiliser au moins des modèles bigrammes de faon à obtenir de meilleurs résultats.

Conclusions et Perspectives

La première partie de cette thèse a été consacrée à la génération automatique des prononciations pour des OOVs et des variantes de prononciation pour les formes de base d'un dictionnaire de reconnaissance. Certaines approches innovantes inspirées de la traduction automatique ont été proposées et des résultats g2p au niveau d'un système état de l'art ont été obtenus sur une baseline difficile. Ensuite, le lexique élargi a été testé dans des expériences de reconnaissance de la parole et des améliorations ont été remarquées sur des dictionnaires de référence ne contenant qu'une seule prononciation. Cependant, l'ajout

d'un grand nombre de variantes entraîne une dégradation de performance ASR. Cela met en évidence le problème bien connu de confusabilité phonémique lorsque de la variation phonémique est ajoutée à un système ASR sans aucune contrainte. Notre intérêt s'est alors orienté dans la direction d'avoir une meilleure compréhension de ces phénomènes de confusabilité et de trouver un moyen de les mesurer et de les contrebalancer.

L'"entropie de prononciation" a alors été définie pour mesurer la confusabilité introduite par le lexique de reconnaissance dans le processus de décodage. Des expériences ont été menées afin d'observer comment cette mesure est influencée lorsque des variantes générées automatiquement sont ajoutés au lexique. Nous avons également mesuré l'influence de l'utilisation de fréquences pour pondérer les prononciations du lexique en la contrastant avec la situation où les variantes sont non pondérées. Nous n'avons cependant pas réussi à trouver une corrélation claire entre cette mesure et le taux d'erreur du système.

L'utilisation de comptes de fréquence est une façon très simpliste d'attribuer des poids aux prononciations et reste limité aux mots qui sont présents dans le corpus d'entraînement. Un moyen plus approprié de choisir des prononciations est d'entraîner leurs poids pour améliorer la performance d'un système ASR. Dans cette thèse, l'apprentissage discriminant a été proposé afin d'entraîner un modèle de confusion des phonèmes qui élargit l'espace de recherche de prononciation lors du décodage ASR. Les méthodes proposées introduisent de la variation phonémique tout en gardant la confusabilité du système à un niveau faible. De plus, la variation supplémentaire est adaptée à un corpus de données particulier et n'est pas statique comme dans la tâche de conversion g2p. Des méthodes d'apprentissage et de décodage basées sur les FST ont été mises en œuvre et des amélioration sur le décodeur FST ont été observées. Il n'est pas évident cependant d'intégrer notre modèle de confusion dans un décodeur non-FST.

Enfin, nous avons élargi l'apprentissage discriminant à la tâche KWS avec l'adoption d'une nouvelle fonction objectif, directement liée à la performance KWS. Il existe un intérêt croissant pour la tâche KWS parce que la quantité de données audio disponibles augmente de façon exponentielle, et qu'il devient indispensable de disposer de moyens efficaces de les fouiller afin de pouvoir en tirer la meilleure utilisation. Dans ce travail, des gains ont été observés sur la baseline lors de l'utilisation d'un modèle de confusion de phonèmes entraîné de façon discriminante pour développer l'index d'un système basé sur des phonèmes.

Nous pensons que la plupart des perspectives de ce travail tournent autour de la construction dynamique d'un lexique de prononciation. Ce fut effectivement le chemin que cette thèse a pris au fur et à mesure que nous avons progressé dans notre connaissance des problèmes et des besoins des modèles de prononciation. Un dictionnaire statique peut donner un très bon rappel des résultats dans l'évaluation d'un convertisseur g2p, mais cela ne se traduira pas toujours par une amélioration de la performance d'un système ASR. L'entraînement spécifique des données du lexique de reconnaissance (en utilisant des données audio) est préférable, comme on le fait traditionnellement pour d'autres parties d'un système ASR. Il est de notre conviction que l'adaptation du lexique basé sur des données va devenir de plus en plus important car nous nous éloignons de la reconnaissance de parole "normée" (Broadcast News), qui est plus ou moins caractérisé par des

locuteurs qui suivent la forme de base de prononciation. Il existe un intérêt croissant pour la reconnaissance de parole plus spontanée ou plus variée en raison des accents que les dictionnaires statiques ne parviennent souvent pas à modéliser correctement.

Pour construire un lexique de prononciation adaptée dynamiquement, nous devons d'abord trouver les bonnes variantes, puis leurs poids afin de leur donner un pouvoir discriminant par rapport aux autres. C'est ce que avons cherché à faire dans le cadre de l'apprentissage discriminant sur les poids des paires de confusion de phonèmes pour capturer la variation de prononciation. En ce qui concerne les caractéristiques phonémiques utilisés, ajouter plus de contexte est une voie de recherche prometteuse. Il pourrait également être bénéfique d'expérimenter avec des traits qui fournissent des informations provenant d'autres sources, par exemple d'intégrer du contexte prosodique ou syntaxique.

Concernant l'apprentissage discriminant proposé dans cette thèse, nous avons l'intention de mener des expériences avec des fonctions objectif dans lesquelles le coût est directement intégré. Il peut également être intéressant d'essayer d'ajouter l'"entropie de prononciation" dans la fonction objectif, ce qui offrirait certaines informations supplémentaires qui pourraient permettre un meilleur contrôle de la confusabilité. Un cadre fondé sur les FST est déjà mis en place, qui permet ces transformations, sans beaucoup de programmation supplémentaire ou de coût de calcul plus élevé. Un de nos objectifs est également d'effectuer l'apprentissage au niveau des mots en conservant les features de confusions phonémiques, mais en permettant l'ajout de features basées sur mots (par ex., une paire de confusion de phonèmes survenant souvent dans un mot particulier). Cela permettra de mieux tenir compte des erreurs les plus courantes du système qui sont difficiles à corriger autrement. Les mêmes modèles pourraient être utilisés pour d'autres tâches sans aucun changement particulier si des données appropriées sont disponibles, par exemple pour la tâche de correction orthographique. Pour la tâche KWS, il y a la possibilité d'intégrer le modèle de confusion proposé dans un système hybride. De cette façon, des gains supplémentaires pourraient être obtenus parce que la méthode proposée serait appliquée aux mots OOVs qui sont ceux qui provoquent très souvent des erreurs. En outre, tant pour le décodage ASR que pour KWS, nous aimerions tester nos modèles sous différentes conditions expérimentales, mieux contrôlées et ciblées sur un problème particulier. Par exemple, nous pourrions voir plus grandes améliorations en travaillant avec de la parole accentuée, où notre modèle de confusion pourrait capturer, par exemple, certaines substitutions communes ou des suppressions phonémiques specifiques à un accent donné.