Alignement automatique et analyse phonétique : comparaison de différents systèmes pour l'analyse du schwa

Audrey Bürki* — Cédric Gendrot —Guillaume Gravier*** — George Linarès**** — Cécile Fougeron****

* *Laboratoire de Psycholinguistique Expérimentale, Université de Genève Genève, Suisse, audrey.buerki@unige.ch*

** *Laboratoire de Phonétique et Phonologie, UMR7018, CNRS-Paris3/Sorbonne Nouvelle, Paris, France, {cedric.gendrot, cecile.fougeron}@univ-paris3.fr*

*** *Institut de Recherche en Informatique et Systèmes Aléatoires, CNRS UMR 6074, Rennes, France, ggravier@irisa.fr*

**** *Laboratoire Informatique d'Avignon, Université d'Avignon et des Pays du Vaucluse, Avignon, France, georges.linares@univ-avignon.fr*

RÉSUMÉ. *L'adéquation des alignements en phonèmes fournis par trois systèmes d'alignement automatique pour l'analyse linguistique du schwa est évaluée par comparaison avec un alignement manuel réalisé par deux juges. Les taux et types d'erreurs sont rapportés ainsi que les facteurs linguistiques ayant une incidence sur ces derniers. Les performances des trois systèmes sont influencées en particulier par la nature des consonnes entourant le schwa et par la durée de celui-ci. Des différences sont néanmoins observées entre les alignements des trois systèmes en fonction de la tâche (détection* vs *placement des frontières) et des caractéristiques structurelles des systèmes. Ces données soulignent l'importance d'un choix réfléchi du système à utiliser et d'une bonne connaissance de ses caractéristiques pour son utilisation à des fins d'analyse phonétique.*

ABSTRACT. *Three automatic alignment systems are compared in their adequacy to account for the vowel schwa as compared to a manual transcription obtained from two judges. Error rates and types are analysed, as well as the linguistic factors involved. The type of surrounding consonants and the duration of schwa influence the decisions of the three systems. Moreover, the systems behave differently, depending on some of their architectural characteristics and on the task (detection vs. boundary determination). These results underline the necessity to have a fair understanding of the characteristics and bias of the alignment system to be used in a phonetic analysis.*

MOTS-CLÉS : *alignement automatique, schwa, corpus, analyse phonétique.*

KEYWORDS: *automatic alignment, schwa, corpus, phonetic analysis.*

TAL. Volume 49 – n° 3/2008, pages 165 à 197

1. Introduction

Si les rapports entre les recherches en traitement automatique de la parole (TAP) et celles effectuées en phonétique ont fluctué au cours des années, l'évolution technologique a suscité récemment un regain d'intérêts communs, qui prend forme autour du traitement de grands corpus de parole continue. En effet, phonéticiens et phonologues cherchent de plus en plus à confronter leurs hypothèses sur des données de parole produite dans des situations plus naturelles et variées qu'en laboratoire, et sur des corpus moins contrôlés et suffisamment importants. Les grands corpus recueillis et utilisés dans le cadre du TAP, ainsi que les instruments automatiques qui y sont développés, constituent de ce fait une mine d'or que beaucoup souhaitent pouvoir exploiter.

Dans ce travail, nous nous intéressons en particulier aux outils automatiques d'alignement en phonèmes et à leur utilisation dans le cadre d'études phonético-phonologiques. L'alignement en phonèmes prend comme source un signal de parole et détermine la succession des phones produits ainsi que leurs frontières. Sa réalisation manuelle étant extrêmement longue et laborieuse, le recours à un alignement automatique permet un gain de temps considérable. Plusieurs corpus alignés automatiquement – ou outils d'alignement – ont récemment été mis à disposition et utilisés par des phonéticiens (voir entre autres (Gendrot et Adda-Decker, 2005 ; Fougeron *et al.,* 2007 ; Kuperman *et al*., 2007)). La question que nous souhaitons aborder concerne la fiabilité d'une telle démarche. Plusieurs points sont ici à considérer. Il existe tout d'abord différents systèmes d'alignement (voir section 2) dont les caractéristiques sont susceptibles de donner lieu à des performances inégales (van Bael *et al*., 2007). Ces systèmes ont par ailleurs, pour nombre d'entre eux, été développés dans le cadre de la reconnaissance de la parole. Or, un fort taux de reconnaissance correcte des mots ne va pas toujours de pair avec un alignement en phonèmes des plus adéquats (Kessens et Strik, 2004). Finalement, on sait l'importance de l'objectif pour lequel un alignement a été réalisé sur les caractéristiques de ce dernier (van Bael *et al*., 2007). Il devient dès lors légitime, voire nécessaire, de s'interroger sur l'adéquation d'une utilisation, à des fins linguistiques, d'alignements qui n'ont pas été effectués dans ce but.

Dans le cadre de la présente étude, les alignements en phonèmes de trois systèmes automatiques sont évalués et comparés, relativement à leur pertinence pour l'étude linguistique d'une voyelle susceptible d'être élidée, le schwa en français. Il s'agit tout d'abord de déterminer dans quelle mesure les segmentations obtenues sont à même de rendre compte de cet objet linguistique, et, par là, d'être utilisées pour des analyses linguistiques fines. Nous tentons également de dégager les régularités et les facteurs qui régissent les décisions des systèmes et leurs erreurs au regard d'un alignement manuel de référence. La pertinence d'une généralisation des données obtenues est évaluée.

2. L'alignement automatique

2.1. *Généralités et paramètres modifiables*

La nécessité de disposer de larges corpus alignés phonétiquement pour le TAP a amené les chercheurs à automatiser l'alignement (Brugnara *et al*., 1993 ; Cucchiarini et Strik, 2003). Différentes méthodes sont utilisées, en fonction notamment de la forme du corpus disponible. La reconnaissance de phones[1] permet d'obtenir un alignement automatique à partir des seuls enregistrements audio. Cependant, sans contraintes, cette reconnaissance donne des résultats insuffisants pour la plupart des applications, avec seulement 50 à 80 % de phones reconnus correctement (Wester *et al*., 2001). Lorsque la transcription orthographique des enregistrements est disponible ou que le corpus est phonétisé, un autre type d'alignement peut être réalisé, communément appelé « alignement forcé ». La transcription orthographique est tout d'abord convertie en une transcription phonétique canonique, par l'accès à un lexique ou par une conversion automatique graphème-phonème. Un certain nombre de variantes de prononciations sont générées sur la base de cette transcription canonique, par l'application de règles phonologiques (Boula de Mareüil et Adda-Decker, 2002), de règles fondées sur les données (Kessens et Strik, 2004), par le moyen d'arbres de décision (Riley *et al*., 1999) ou encore par correction manuelle du lexique. On obtient au final un graphe des prononciations possibles pour une phrase donnée. L'alignement phonétique forcé entre le signal et la phrase s'apparente à la recherche du meilleur chemin dans le graphe de prononciations étant donné une représentation du signal, typiquement des coefficients cepstraux. Les outils développés pour la reconnaissance de mots sont alors très souvent sollicités (Kessens et Strik, 2001), en particulier les modèles statistiques de la parole dont les modèles de Markov cachés (HMM). Dans ce dernier cas, à partir de modèles élémentaires pour chaque phone et du graphe de prononciations, l'algorithme de Viterbi permet de calculer de manière efficace le chemin optimal dans le graphe de prononciations étant donné le signal analysé. Le résultat d'un tel procédé d'alignement est, d'une part, l'identification des phonèmes prononcés et, d'autre part, la localisation temporelle de leurs frontières. Plusieurs paramètres sont déterminants, tels que la qualité des phonétisations, le nombre de variantes listées dans le dictionnaire de prononciations et leur pertinence (voir notamment Strik et Cucchiarini, 1999). Par ailleurs, l'estimation des vraisemblances dans les HMM est réalisée par des mélanges de gaussiennes dont la précision et la complexité sont variables et fonction à la fois des corpus et des stratégies mises en œuvre pour leur apprentissage.

Une étude de (Kessens et Strik, 2004) donne un aperçu de quelques-uns des paramètres qu'il est possible de modifier afin d'optimiser la qualité d'un alignement

1. Le terme de phone traduit la réalisation d'un phonème. En TAP, la relation entre phone et phonème n'est pas toujours univoque : des réalisations de différents phonèmes sont parfois regroupées sous une même étiquette. C'est par exemple souvent le cas pour le schwa et /œ/.

automatique. Cette dernière est ici entendue en termes de similarités avec un alignement manuel de référence sur la présence/absence d'un ensemble de phones susceptibles d'être élidés ou insérés suite à l'application d'une règle phonologique[2]. Les paramètres renvoient à certaines des propriétés des HMM : unités de base dépendantes ou non du contexte phonétique, structure des modèles (topologie) et degré de contamination. Les HMM dépendants du contexte prennent en compte le contexte dans lequel le phone a été produit. Étant mieux à même de modéliser les effets de contexte tels que la direction des transitions et la coarticulation, ils permettent en général un taux de reconnaissance correcte des mots plus élevé. En ce qui concerne les performances de l'aligneur cependant, les auteurs observent un taux d'erreurs plus important pour les HMM dépendants du contexte. Ce résultat s'explique selon eux par la plus grande contamination de ces derniers, davantage biaisés en faveur du corpus canonique sur lequel ils ont été entraînés que les HMM indépendants du contexte. La topologie des HMM (notion qui réfère au nombre d'états d'un HMM) peut également être modifiée. Le modèle de durée implicite d'un HMM dépend notamment de sa topologie, c'est-à-dire du nombre d'états qui le composent et des transitions entre ces états. Traditionnellement, les modèles gauche-droite à 3 états imposent une durée phonémique minimale de 30 ms (10 ms par état). Les auteurs observent qu'en autorisant des chemins minimaux de 20 ms, les performances d'alignement s'améliorent sensiblement. Les options prises lors de l'apprentissage des modèles acoustiques et notamment concernant la modélisation de la variation à ce niveau sont également déterminantes pour la qualité de l'alignement automatique. De meilleurs résultats sont ici obtenus en réduisant la « contamination » des HMM. Cela est rendu possible en entraînant les modèles de phones sur du texte lu ou en ajoutant aux HMM de base des HMM entraînés sur un corpus dans lequel certaines formes de variation sont modélisées. Dans une étude antérieure, (Kessens et Strik, 2001) ont également examiné l'effet de la modification de la résolution acoustique des HMM (nombre de gaussiennes par état) sur la qualité de l'alignement sans que des tendances claires puissent être dégagées.

Ces quelques résultats donnent un aperçu des différences qui peuvent exister parmi les systèmes d'alignement ayant recours à l'alignement forcé et donc sous-tendus par une démarche similaire. Cette liste n'est pas exhaustive, d'autres paramètres sont modifiables et les options qu'il est possible de leur faire emprunter sont multiples. Il s'agit par ailleurs de souligner que ces différents paramètres interagissent généralement sur les performances de l'aligneur, et que leur combinaison doit également être considérée (Kessens et Strik, 2004 ; Adda-Decker et Lamel, 2000).

Nous pouvons déduire de ce survol l'existence à la fois de grandes similarités entre les systèmes d'alignement utilisés actuellement et de nombreux points de divergence. Une analyse des implications de ce double constat sur les alignements

2. Il s'agit d'un corpus en néerlandais. Les règles concernées sont : l'effacement du /n/, du /ʀ/, du /t/ et du /ə/ et l'insertion du /ə/.

effectués par ces différents systèmes pourrait déboucher sur des conclusions non triviales pour l'utilisation de ces alignements à des fins linguistiques. En particulier, il serait intéressant de savoir dans quelle mesure les caractéristiques communes à ces différents systèmes conduisent à des similarités dans leurs alignements. Si de nombreuses généralités peuvent être dégagées, la nécessité pour le linguiste d'évaluer tout nouveau système s'en trouvera limitée. Si au contraire l'ajustement des différents paramètres joue un rôle prépondérant dans la qualité de l'alignement, les résultats de l'évaluation d'un système donné ne pourront être généralisés à l'emploi d'autres systèmes. Bien que ce dernier cas de figure paraisse présenter un inconvénient majeur pour le linguiste, de nouvelles opportunités se dessinent également. Il devient en effet possible d'envisager, à partir des outils dont nous disposons actuellement, de paramétrer les systèmes existants pour les besoins spécifiques d'analyses linguistiques.

2.2. *Évaluation des alignements automatiques : questions méthodologiques*

L'évaluation des résultats d'un alignement automatique soulève plusieurs problèmes (Wester, 2001 ; Wesenik et Kipp, 1996). La difficulté majeure est liée à la nécessité de pouvoir comparer le résultat obtenu avec un alignement de référence. Un alignement manuel est en général utilisé à ces fins. Or, plusieurs études ont comparé les performances de transcripteurs manuels entre elles et montrent des disparités parfois importantes. Dans l'étude de (Wesenick et Kipp, 1996), le pourcentage de phones étiquetés de manière similaire par les deux transcripteurs est de 95 % (contre 88 % entre alignement manuel et alignement automatique). Par ailleurs 87 % des frontières communes sont situées dans un intervalle inférieur à 10 ms, 96% dans un intervalle inférieur à 20 ms (contre 61 % et 84 % respectivement pour la comparaison manuel/automatique). Lorsque le nombre de transcripteurs augmente, les disparités sont plus flagrantes. (Pitt *et al.*, 2005) ont comparé la transcription de quatre transcripteurs entraînés. 64 % seulement des 2 159 phones reçoivent un étiquetage unanime, 80 % d'entre eux font l'objet d'un accord entre une des six paires de juges au moins. En ce qui concerne la présence même du segment, lorsqu'un des transcripteurs considère un phone comme étant présent, dans 86 % des cas seulement les autres transcripteurs l'ont également segmenté. Par ailleurs, ces auteurs relèvent une déviation moyenne de 16 ms en ce qui concerne le placement des frontières.

Si les alignements de deux transcripteurs manuels sont toujours plus proches que lorsque la comparaison est effectuée entre un alignement manuel et un alignement automatique, les disparités existantes entre alignements manuels mettent en question la légitimité du recours à une référence établie manuellement. Deux options ont été proposées dans la littérature pour établir une référence acceptable. La première consiste à ne considérer dans les évaluations que le matériel linguistique sur lequel tous les transcripteurs s'accordent. C'est la méthode du consensus, décrite entre

autres par (Shriberg *et al.*, 1984). La seconde option consiste à construire une transcription de référence fondée sur le vote majoritaire (Wester *et al*., 2001, voir également (Kessens et Strik, 2004). Ces deux propositions présentent cependant des inconvénients majeurs lorsque l'on s'intéresse aux corpus à des fins d'analyses linguistiques. Les désaccords entre transcripteurs peuvent être porteurs d'informations linguistiques pertinentes. Par exemple, dans une étude antérieure sur le schwa en français (Bürki *et al.*, 2007) nous avons montré qu'il existe un certain nombre d'occurrences pour lesquelles il est difficile de trancher en faveur de la présence/absence de la voyelle. L'étude de ces cas ambigus suggère qu'ils concernent certains contextes segmentaux uniquement, observation dont les implications théoriques sont importantes, notamment en lien avec le débat autour de la nature du processus en jeu dans l'alternance. Par ailleurs, l'éradication des « cas limites » est susceptible d'orienter sensiblement les résultats. Ces dernières années, la question de la nature binaire ou graduelle de différents processus phonétiques et/ou phonologiques a été soulevée. L'accès à de grands corpus de parole continue pourrait permettre d'éclaircir ces questions dans les années à venir. Or, l'appréhension de processus graduels ne peut se faire si ces corpus sont dépouillés des « cas limites », forcément litigieux.

2.3. *Évaluation des alignements automatiques : quelques résultats*

La fiabilité des alignements automatiques, entendue en termes de similarités/écarts avec un alignement manuel de référence, a été évaluée à plusieurs reprises dans la littérature, généralement dans le cadre du TAP. Des mesures d'accord entre alignement manuel de référence et automatique sont rapportées, elles concernent le plus souvent les scores de détection de phones. La valeur du coefficient de Kappa est fréquemment utilisée pour rendre compte de la fiabilité d'un alignement. Elle oscille par exemple entre 0,44 et 0,55 dans l'étude de (Kessens et Strik, 2004) mentionnée ci-dessus. (Van Bael *et al*., 2007) proposent, quant à eux, une mesure globale de désaccord, regroupant différents types d'erreurs : insertions (ajout de phonèmes non présents dans l'alignement de référence), effacements (phone présent dans l'alignement de référence mais non détecté par l'alignement automatique) et substitutions (étiquette attribuée par l'alignement automatique différente, exemple : un /p/ dans l'alignement de référence peut être étiqueté /b/ dans l'alignement automatique). Les pourcentages de désaccord globaux les plus élevés sont de 6 % pour de la parole lue et de 5 % pour des conversations téléphoniques. Plus rarement, des mesures du décalage entre frontières manuelles et automatiques sont fournies. (Auran et Bouzon, 2003) et (Sjölander, 2003), rapportent respectivement 70 % et 85 % des frontières de l'alignement automatique tombant dans un intervalle inférieur à 20 ms au regard de l'alignement manuel.

Si ces données quantitatives sont intéressantes et nécessaires, elles ne nous disent rien de la qualité des différences observées entre alignements manuels et automatiques. (Pitt *et al*., 2005) soulignent que lorsqu'un corpus aligné manuellement est utilisé pour l'étude des variantes de prononciations, il est important de comprendre dans quelle mesure et comment les choix du transcripteur contribuent à la variation mise en évidence. Si ces auteurs parlent ici d'alignement manuel, cette réflexion s'applique également, voire davantage à l'alignement automatique, et ce quel que soit l'objet d'analyse linguistique. Il est en effet crucial, dans l'optique d'une analyse linguistique, de savoir si les différences entre l'alignement manuel et l'alignement automatique concernent l'ensemble des phonèmes et des contextes de manière égale, ainsi que de pouvoir énumérer les facteurs qui influent sur ces différences. Ces informations doivent permettre de départager ce qui relève de phénomènes linguistiques de ce qui résulte des choix de l'aligneur, sous peine de proposer une interprétation linguistique inadéquate des données. Plusieurs études fournissent quelques éléments dans cette direction. (Wesenick et Kipp, 1996) mettent en évidence certaines similarités entre alignement automatique et alignement manuel sur un corpus allemand. Les types de confusions concernant les segments consonantiques sont similaires (/d/ et /t/ sont par exemple souvent confondus). Par ailleurs, les frontières difficiles à placer pour les humains sont également source de difficultés pour la machine. (Binnenpoorte *et al*., 2004) détaillent les types d'erreurs d'alignement (désaccords entre alignement manuel et automatique) dans un corpus néerlandais – substitutions, effacements et insertions de segments – et leurs pourcentages (taux d'erreurs global de 22 %), qu'ils relient aux processus phonétiques en jeu. Les types de segments sont considérés. Concernant les substitutions de consonnes, les confusions les plus fréquentes impliquent des phonèmes ne se distinguant que par un seul trait articulatoire. Le schwa est quant à lui le phonème le plus fréquemment impliqué dans les erreurs touchant les voyelles. La majorité des effacements sont liés aux phonèmes /ə/, /ʀ/, /d/ et /n/ et les insertions impliquent le plus souvent les phonèmes /ə/, /ʀ/, /t/ et /n/. (Nguyen et Espesser, 2004) évaluent quant à eux la précision avec laquelle l'aligneur se montre capable de localiser l'emplacement des frontières entre phonèmes dans un corpus français. La congruence entre les alignements manuel et automatique est meilleure en début qu'en fin de voyelle, la durée attribuée aux segments par l'aligneur automatique est plus courte. Leurs données suggèrent par ailleurs que les écarts entre étiquettes manuelles et automatiques sont conditionnés par le contexte phonétique dans lequel la voyelle se présente. Finalement, si le milieu de la voyelle est localisé avec une bonne précision (écart inférieur à 20 ms) dans 75 % des cas, la précision varie en fonction de la position de la voyelle dans le mot et du contexte segmental droit. La semi-consonne /j/ et les consonnes /ʀ/ et /z/ donnent en particulier de mauvais résultats. Ces différents résultats suggèrent que les écarts entre alignements manuel et automatique ne concernent pas de manière univoque tous les phonèmes ni les contextes, que le type d'erreur considéré ou la mesure rapportée influencent le résultat et que différents facteurs semblent gouverner ces écarts.

3. Le cas d'étude : le schwa en français

La nécessité de rapporter l'évaluation d'un alignement automatique aux applications auxquelles cet alignement est destiné a déjà été mentionnée. Dans cette étude, plusieurs systèmes d'alignement automatique sont évalués dans leur capacité à rendre compte d'un objet linguistique particulier en français : la voyelle schwa.

La voyelle schwa (ou « e » muet) a la particularité phonologique d'alterner avec zéro (Malécot, 1977 ; Côté et Morrison, 2007). Autrement dit, un mot comportant un schwa peut être prononcé avec ce dernier (ex. : [səmɛn]) ou sans (ex. : [smɛn]). Les raisons de ce choix sont multiples. Si le schwa a largement occupé la littérature linguistique depuis plusieurs décennies, il est aujourd'hui encore difficile d'en brosser un tableau consensuel. Nombreuses sont les questions non encore résolues à son propos et pour lesquelles le corpus constitué fournira une base de données à exploiter. Par ailleurs, le schwa pose des problèmes non négligeables aux systèmes d'alignement utilisés pour la synthèse de la parole (Lanchantin *et al.*, 2008), ainsi qu'aux systèmes de reconnaissance. (Adda-Decker, 2007) relève que 5 % des erreurs de reconnaissance impliquent des omissions, insertions ou confusions liées au schwa. Mieux cerner sa gestion par les systèmes d'alignement pourrait donc être profitable à la recherche en TAP. De plus, comprendre comment les différents systèmes d'alignement se comportent vis-à-vis de l'élision du schwa pourrait nous offrir un premier aperçu des problèmes que peut poser le phénomène plus large de réduction/effacement de segments dans la parole continue en français et dans d'autres langues (Wester *et al*., 2001 ; van Bael *et al*., 2007 ; Binnenpoorte *et al.,* 2004 ; Adda-Decker *et al*., 2000), effacements difficiles à gérer par les systèmes de reconnaissance (Strik *et al*., 2006).

4. Méthode

4.1. *Matériel linguistique*

Un répertoire de mots contenant un schwa a été constitué par le regroupement de différentes bases lexicales existantes : Lexique (New *et al.,* 2001), Brulex (Content *et al.,* 1990), IlPho (Boula de Mareüil *et al.,* 2000), complété par un certain nombre de mots issus du dictionnaire « Le Grand Robert ». Après élimination des entités nommées, mots composés et schwas frontières de clitiques, le nombre de mots se montait à 18 553. Le corpus ESTER (Galliano *et al.,* 2005) a ensuite été consulté afin d'y rechercher les occurrences des mots du répertoire et en extraire les fichiers sons correspondants. La partie du corpus ESTER concernée, commune aux différents alignements utilisés dans cette étude inclut 24 heures de parole radiophonique issues de radios francophones et produites par 574 locuteurs. Il s'agit essentiellement de parole planifiée, avec cependant quelques heures de parole spontanée. 22 773 occurrences pour un total de 583 mots contenant un schwa ont

été extraites. La présente analyse ne s'intéressant qu'aux schwas alternants[3], il était nécessaire de classer les mots en ces termes. Une première classification, réalisée à partir de l'intuition de trois juges, a révélé de nombreuses disparités entre les alternances considérées comme optionnelles, obligatoires, ou même interdites. Un critère objectif a de ce fait été appliqué : seuls les mots apparaissant de manière alternante dans le corpus tel que le proposait l'alignement automatique ont été retenus pour l'analyse (mots ayant été segmentés avec et sans schwa, par exemple le mot « semaine » segmenté 84 fois [səmɛn] et 36 fois [smɛn]). 5 016 occurrences ont été soumises à une analyse auditive et acoustique. 230 séquences impropres à l'analyse (5 % des données) ont été exclues : 135 occurrences produites par des locuteurs jugés non francophones sur la base de leur accent (22 locuteurs) et 95 séquences inintelligibles ou correspondant à des erreurs de reconnaissance (voir (Bürki *et al.*, 2007) pour davantage de détails sur la procédure). Une première correction manuelle de la segmentation automatique (effectuée par un juge et entérinée par un second) a permis d'éliminer 67 mots (479 occurrences) ayant été produits de manière non variable (uniquement avec ou uniquement sans schwa) dans le corpus. Au final, 4 307 occurrences ont été retenues pour l'analyse[4].

4.2. *Alignement manuel*

Afin d'établir un alignement de référence, nous avons opté pour une démarche qui tienne compte des différentes considérations évoquées dans la section 2.2. Un premier juge a corrigé la segmentation automatique du système de l'IRISA (Institut de Recherche en Informatique et Systèmes Aléatoires) fondé sur des monophones (voir section 4.3), l'objectif étant de pouvoir ensuite utiliser cet alignement pour l'étude du schwa. Un second juge a fait de même sur 47 % des occurrences, en ayant accès à la segmentation du premier juge. Un degré d'accord a été calculé afin de pouvoir ensuite estimer les résultats des systèmes automatiques. En effet, afin de déterminer si un alignement phonétique obtenu automatiquement est satisfaisant ou non, il est nécessaire de disposer d'une ligne de base. Plusieurs auteurs utilisent à cet effet le degré d'accord entre deux alignements manuels (Cucchiarini et Strik, 2003). Si le degré d'accord entre alignement automatique et alignement manuel est similaire à celui observé entre deux alignements manuels, le système d'alignement est jugé satisfaisant.

---

3. Suivant la définition de (Côté et Morrison, 2007) nous ne considérons ici comme schwa que les voyelles alternantes (qualifiées par d'autres de schwas « optionnels »). Cette définition exclut les schwas obligatoires ou jamais produits.

4. Tous les traitements de fichiers sons mentionnés dans cet article on été effectués avec Praat (Boersma et Weenink, 2007).

4.2.1. *Critères/consignes de transcription*

À partir de l'alignement automatique de l'IRISA fondé sur des monophones plusieurs modifications ont été effectuées par chacun des juges. Les schwas non détectés ont été ajoutés, les schwas détectés par erreur (insertions) éliminés. Un critère uniforme a été appliqué : un schwa a été considéré comme réalisé en présence à la fois de périodicité dans le signal et d'une structure formantique (voir (Bürki *et al*., 2007), pour davantage de détails sur le choix de ces critères et ses implications[5]). La correction des estimations de durée a ensuite été effectuée sur la base de l'apparition/disparition de périodicité sur le signal acoustique et d'un deuxième formant sur le spectrogramme. Différentes analyses[6] ont été entreprises afin d'évaluer le degré d'uniformité entre les deux alignements manuels. Ces dernières ont porté sur la présence de la voyelle, l'estimation de sa durée et le placement des frontières.

4.2.2. *Mesure de la fiabilité interjuges sur la présence/absence du schwa*

Deux mesures de fiabilité interjuges pour les jugements dichotomiques sont généralement rapportées dans la littérature, chacune présentant ses intérêts et ses inconvénients (Stemler, 2004). Ces deux mesures ont été utilisées ici afin de comparer les alignements manuels des deux transcripteurs en termes de présence/absence du schwa. Des pourcentages d'accord ont tout d'abord été calculés. L'accord global (nombre d'occurrences pour lesquelles le jugement des deux juges est identique sur nombre total de jugements) équivaut à 99 %. L'accord est de 98 % sur les schwas présents et de 99 % sur les schwas absents. La relation entre les deux jugements est significative ($\chi^2(1) = 1914{,}9$, $p < 0{,}0001$) et de force importante ($\varphi = 0{,}977$). La seconde mesure souvent utilisée pour mesurer le degré d'accord entre deux juges sur des données catégorielles est le coefficient de Kappa (Cohen, 1968). Ce dernier permet de corriger les pourcentages en tenant compte du hasard et rend possible une comparaison entre des valeurs issues de différentes conditions (voir cependant (Stemler, 2004)).

5. Dans (Bürki *et al*., 2007) nous avons observé que certaines occurrences présentaient des traces acoustiques d'un élément vocalique moins saillantes que celles considérées ici comme témoignant de la présence d'un schwa. Dans le cadre de la présente étude ces traces n'ont pas été attribuées au schwa pour plusieurs raisons. D'une part, des études ultérieures sont nécessaires afin de départager ce qui relève du schwa de ce qui relève de transitions articulatoires entre les consonnes. D'autre part, le présent travail n'a pas pour objectif de juger les systèmes d'alignement automatique sur un tel niveau de précision. Étant donné la tendance des systèmes automatiques à ne pas repérer les schwas présentant des indices acoustiques saillants, leurs performances seraient sans aucun doute plus faibles si elles devaient être jugées sur ce niveau de détails.

6. Toutes les analyses statistiques présentées dans cet article ont été effectuées avec le logiciel SPSS version 15.0 (SPSS Inc., Chicago IL).

$$k = \frac{Po - Pc}{100 - Pc} = 0{,}98$$

où : Po = % d'accord observé et
Pc = % d'accord dû au hasard.

D'après (Landis et Koch, 1967), un k égal à 0 indique un accord dû au hasard, un k égal à 1 indique un accord parfait. L'accord est jugé faible entre 0 et 0,2, acceptable entre 0,21 et 0,4, modéré entre 0,41 et 0,6, substantiel de 0,61 à 0,8 et presque parfait de 0,81 à 1. Le coefficient obtenu ici (0,98), particulièrement élevé, témoigne d'un accord « presque parfait ». Pour comparaison, le taux d'accord entre quatre transcripteurs sur la présence du schwa rapporté par (Pitt *et al*., 2005) est de 86 %. (Wester *et al*., 2001) rapportent, quant à eux, un coefficient de kappa égal à 0,8 entre deux alignements manuels en ce qui concerne l'insertion du schwa et à 0,34 en ce qui concerne l'effacement de ce dernier.

4.2.3. *Fiabilité interjuges sur l'estimation de la durée du schwa*

La durée moyenne des 1 420 schwas jugés présents par les deux juges a été calculée. Elle est de 52 ms, pour le premier et de 53 ms pour le second. Le coefficient de corrélation entre les estimations des deux juges est élevé ($r = 0{,}966$, $n = 1\,420$, $p < 0{,}01$) et ce malgré une différence de durée significative ($t(1\,419) = 8{,}38$, $p < 0{,}001$).

4.2.4. *Fiabilité interjuges sur le placement des frontières*

95 % des frontières placées par les deux juges tombent dans un intervalle de 0 à 10 ms et 99 % tombent dans un intervalle de 0 à 20 ms. Ces valeurs sont plus hautes que celles rapportées notamment par (Wesenick et Kipp, 1996), pour des segments consonantiques (87 % des frontières dans un intervalle de 10 ms et 96 % dans un intervalle de 20 ms). La déviation moyenne est de 9 ms pour le début de la voyelle et de 6 ms pour sa fin. Ces valeurs sont plus basses que ce qui a été obtenu par (Pitt *et al*., 2005) dont l'évaluation est fondée sur l'alignement de quatre transcripteurs et concerne l'ensemble des segments vocaliques et consonantiques (écart moyen de 16 ms). Par ailleurs, le second juge tend à placer ses frontières de début de voyelle plus à gauche alors que les frontières de fin de voyelle sont décalées vers la droite.

Dans leur ensemble, ces résultats révèlent un degré d'accord extrêmement important entre les deux alignements manuels, généralement plus haut que ce qui est rapporté dans la littérature. Les comparaisons restent cependant difficiles, étant donné des différences importantes de matériel et de méthodologie. Le haut degré d'accord obtenu ici est probablement à rapporter à la méthode hybride utilisée, entre des alignements indépendants (qui donneraient des résultats plus faibles) et un consensus. Il nous permet de recourir, pour l'évaluation des alignements automatiques, à l'alignement du premier juge, portant sur la totalité des occurrences.

4.3. *Systèmes d'alignement automatique évalués*

Trois systèmes d'alignement sont évalués dans la présente étude, tous développés dans le cadre de la reconnaissance de la parole continue grand vocabulaire. Un système a été développé au LIA (Laboratoire Informatique d'Avignon) et les deux autres à l'IRISA. Les deux systèmes de l'IRISA, que l'on appellera « IRISA monophones » et « IRISA triphones », sont en fait deux variantes d'un même système. Les trois systèmes d'alignement ont recours à des HMM. Ils se distinguent en revanche sur deux points : le dictionnaire de prononciations et les modèles acoustiques utilisés.

Pour les trois systèmes, les dictionnaires de prononciations ont été construits à partir de phonétiseurs automatiques et corrigés à la main. Pour un mot donné, les différentes prononciations, appelées variantes, sont considérées comme équiprobables et ne dépendent pas du contexte syntaxique dans lequel un mot est prononcé. Dans les deux systèmes de l'IRISA, les prononciations ont été établies à partir du dictionnaire ILPho (Boula de Mareüil *et al*., 2000), fondé sur des règles de phonétisation. Seules les variantes de prononciations les plus probables ont été conservées, avec un nombre moyen de variantes de prononciations par mot de 1,8. Le système du LIA, quant à lui, utilise le phonétiseur automatique à base de règles LIA_PHON (Béchet, 2001) dont les sorties ont été corrigées manuellement, pour générer une moyenne de 4,54 variantes de prononciations par mot.

Les modèles acoustiques utilisés par les systèmes du LIA et de l'IRISA ont été estimés sur le corpus d'apprentissage de la base ESTER (Galliano *et al*., 2005), composé de 80 heures d'émissions radiophoniques. Ces modèles diffèrent par leur structure, leur prise en compte ou non du contexte phonétique et leur taille (nombre de densités gaussiennes utilisées). Les deux systèmes de l'IRISA utilisent un ensemble de 35 phonèmes représentés par des HMM à 3 états qui peuvent être, selon le système, soit indépendants du contexte phonétique (système « IRISA monophones »), soit dépendants du contexte phonétique et du sexe du locuteur (système « IRISA triphones »). Dans les deux cas, la structure des modèles utilisés impose une durée minimale des phonèmes de 30 ms. Les modèles indépendants du contexte possèdent au total 114 états modélisés chacun par 128 gaussiennes, soit un total de 14 592 gaussiennes. Les modèles dépendants du contexte possèdent, quant à eux, un total de 8 140 états correspondant à 260 480 gaussiennes, pour les voix d'hommes comme pour les voix de femmes. Les modèles du système d'alignement du LIA comportent 3 400 états, soit environ 230 000 gaussiennes. Ils sont dépendants du contexte phonétique à l'intérieur des mots, mais indépendants de ce dernier en frontière de mots. Par ailleurs, la structure des modèles n'impose pas de contrainte de durée minimale aux phonèmes.

4.4. *Évaluation des alignements automatiques*

Chaque système d'alignement automatique a été évalué en termes de présence/absence de la voyelle, d'estimation de la durée et du placement des frontières au regard de l'alignement effectué par le premier juge. Outre les taux et types d'erreurs, l'influence du contexte consonantique et de facteurs acoustiques intrinsèques au schwa a été mesurée pour chacun des systèmes.

5. Résultats

En raison d'écarts temporels trop importants entre alignement manuel et alignement automatique, un certain nombre d'occurrences n'ont pas pu être considérées dans les analyses. Les comparaisons de chaque système avec l'alignement manuel ne se font donc pas toujours sur le même nombre d'occurrences : 4 287 occurrences pour la comparaison des deux systèmes de l'IRISA avec l'alignement manuel (qu'on nommera « Manuel 1 ») et 4 130 occurrences pour la comparaison du système du LIA avec l'alignement manuel (« Manuel 2 »).

5.1. *Présence/absence du schwa*

5.1.1. *Accord, taux et types d'erreurs*

De manière similaire à ce qui a été entrepris pour la comparaison entre les deux alignements manuels, plusieurs mesures de fiabilité ont été calculées afin de déterminer le degré d'accord entre chacun des alignements automatiques et l'alignement manuel de référence. Le tableau 1 ci-dessous présente les pourcentages d'accord globaux, le coefficient de Kappa, ainsi que le résultat du Chi2 de Pearson pour chacune des comparaisons. L'alignement du LIA diffère davantage de l'alignement manuel de référence que les alignements du système de l'IRISA. Le coefficient de Kappa révèle un accord substantiel (0,78) sur l'échelle proposée par (Landis et Koch, 1967) pour le système du LIA, alors qu'il est « presque parfait » pour les deux alignements du système de l'IRISA.

LIA (n = 4 130)			IRISA monophones (n = 4 287)			IRISA triphones (n = 4 287)		
%	k	χ^2	%	k	χ^2	%	k	χ^2
90,3	0,78	2 528,7 p < 0,0001	92,5	0,83	3 011,8 p < 0,0001	93,4	0,85	3 094,4 p < 0,0001

Tableau 1. *Degré d'accord entre alignement manuel et alignement automatique* : *pourcentage d'accord global (%), coefficient de Kappa (k) et valeur du Chi2 (χ^2)*

Deux types d'erreurs interviennent dans les divergences entre alignement manuel et automatique. Des schwas absents peuvent être segmentés par le système automatique (insertions), des schwas présents peuvent ne pas être détectés. Une analyse distincte de ces erreurs est présentée ci-dessous.

5.1.2. *Insertions : taux et facteurs*

Les trois systèmes se distinguent par le taux d'insertion de schwas dans leurs alignements (nombre d'occurrences alignées avec un schwa sur nombre total d'occurrences sans schwa dans l'alignement manuel). L'alignement du LIA présente le plus fort taux d'insertion (9 %, n = 1 220), suivi de l'alignement IRISA triphones (5 %, n = 1 250) et de l'alignement IRISA monophones (2 %, n = 1 250). Une analyse de régression logistique binomiale confirme l'impact du système sur le taux d'insertion (Chi2 Omnibus : $\chi^2(2) = 46{,}85$, $p < 0{,}0001$). La probabilité qu'un schwa soit inséré est plus faible pour le système IRISA monophones que pour le système IRISA triphones ($z = 3{,}05$, $p < 0{,}01$) ou pour celui du LIA ($z = 6{,}19$, $p < 0{,}0001$). Elle est également plus faible pour le système IRISA triphones que pour le système du LIA ($z = 3{,}67$, $p < 0{,}0001$).

Une analyse du rôle de la nature des consonnes entourant le schwa sur le taux d'erreurs a été conduite. Plusieurs raisons nous ont amenés à focaliser notre attention sur l'influence du mode d'articulation et de la sonorité des consonnes. Durant la phase de correction manuelle de l'alignement, des observations ont pu être faites, allant dans le sens d'une implication de la sonorité des consonnes dans les erreurs de détection. S'agissant du mode d'articulation, son intervention dans d'autres types d'erreurs (estimation de la durée notamment) nous a enjoints à considérer également ce facteur[7]. Par ailleurs des impératifs liés à l'interprétation des données nous ont incités à ne pas analyser chaque consonne séparément. Outre la complexité de lecture des résultats inhérente à une telle démarche, la faible représentation de certaines consonnes aurait rendu les validations statistiques difficiles. Finalement, l'importance de la sonorité et du mode d'articulation des

7. N'ayant pas d'hypothèses particulières liées au rôle potentiel du lieu d'articulation, nous n'avons pas inclus ce trait dans nos analyses.

consonnes est mentionnée par (Wesenick et Kipp, 1996). Ces auteurs soulignent leur influence sur la facilité avec laquelle des suites de consonnes peuvent être segmentées.

Une analyse de régression logistique binomiale a été conduite afin d'évaluer l'influence de l'entourage consonantique (sonorité et mode d'articulation) sur la catégorie de détection (correct *vs* insertion). Pour le système du LIA (n = 1 220), le modèle statistique complet prédit de manière significative la détection (Chi2 Omnibus : $\chi^2(8) = 99{,}8,\ p < 0{,}0001$) et explique entre 8 % (pseudo R^2 de Cox et Snell) et 18 % (pseudo R^2 de Nagelkerke) de la variance. À gauche, une consonne voisée donne lieu à davantage d'insertions qu'une consonne sonante, qui a son tour génère davantage d'insertions qu'une consonne sourde. Concernant le mode d'articulation, le plus grand nombre d'insertions est observé après une consonne nasale, le plus faible après une consonne fricative. À droite, une consonne nasale génère moins d'erreurs qu'une consonne liquide ou fricative.

En ce qui concerne l'alignement du système IRISA monophones (n = 1 250), le modèle complet rend compte des données de manière significative (Chi2 Omnibus : $\chi^2(8) = 23{,}4,\ p < 0{,}01$) et explique entre 2 et 10 % de la variance. À gauche, une consonne fricative génère davantage d'erreurs qu'une consonne liquide ou occlusive. À droite, une consonne nasale donne lieu à davantage d'erreurs qu'une consonne fricative ou liquide, une consonne sonante à davantage d'erreurs qu'une consonne sourde.

Finalement, pour le système IRISA triphones (n = 1 250), le modèle rend compte des données de manière significative (Chi2 Omnibus : $\chi^2(8) = 31{,}2,\ p < 0{,}0001$) et explique entre 3 et 8 % de la variance. À gauche, une consonne voisée donne lieu à davantage d'erreurs qu'une consonne sourde, à droite, une consonne voisée génère davantage d'erreurs qu'une sonante. Le tableau 2 résume l'influence des propriétés des consonnes sur le taux d'insertion pour chacun des systèmes. Nous remarquons en particulier des influences contraires en ce qui concerne le mode d'articulation pour le système du LIA et celui de l'IRISA monophones. Relevons finalement la faible proportion des données expliquées par l'entourage consonantique, notamment dans l'alignement « monophones » de l'IRISA.

Système d'alignement	Consonne gauche	Consonne droite
LIA	voisée > sonante > sourde nasale > occlusive, liquide > fricative	liquide, fricative > nasale
IRISA monophones	fricative > occlusive, liquide	nasale > liquide, fricative sonante > sourde
IRISA triphones	voisée > sourde	voisée > sonante

Tableau 2. *Influences des propriétés des consonnes environnantes sur le taux d'insertion pour chacun des systèmes d'alignement (les contextes à gauche du signe « > » génèrent davantage d'insertions que les contextes à sa droite)*

5.1.3. *Non-détection du schwa : taux et facteurs*

Les trois systèmes se distinguent également par le taux de détection correcte/incorrecte pour les schwas présents dans l'alignement manuel de référence. Le système du LIA présente le plus grand taux de non-détection (10,4 %, n = 2 910) suivi de l'alignement « monophones » de l'IRISA (9,6 %, n = 3 037), puis de l'alignement « triphones » de l'IRISA (8 %, n = 3 037). Une analyse de régression logistique binomiale confirme l'impact du système sur le taux de non-détection (Chi2 Omnibus : $\chi^2(2) = 16{,}3$, $p < 0{,}001$). La probabilité qu'un schwa ne soit pas détecté est plus faible pour le système IRISA triphones que pour le système IRISA monophones ($z = 2{,}93$, $p < 0{,}01$) et que pour le système du LIA ($z = 3{,}87$, $p = 0{,}0001$).

Une analyse de régression logistique binomiale a été conduite pour chaque système afin d'estimer la contribution de différents facteurs sur l'adéquation de la détection pour les schwas présents. Aux caractéristiques des consonnes entourant le schwa a été ajoutée la durée du schwa telle qu'elle a été obtenue manuellement. Pour l'alignement du LIA, le modèle rend compte des données de manière significative (Chi2 Omnibus : $\chi^2(9) = 384{,}3$, $p < 0{,}0001$), il explique entre 12 et 26 % de la variance. Le facteur le plus important est la durée estimée manuellement : plus elle augmente, plus le taux de non-détection diminue. La sonorité des consonnes joue également un rôle. À gauche, comme à droite, les consonnes sonantes donnent lieu à davantage de non-détections que les consonnes sourdes ou voisées, les consonnes voisées à davantage de non-détections que les sourdes. Concernant le mode d'articulation, les consonnes nasales génèrent peu de non-détections par rapport aux consonnes gauches liquides ou fricatives.

Concernant l'alignement IRISA monophones, le modèle prédit les données de manière significative (Chi2 Omnibus : $\chi^2(9) = 443{,}4$, $p < 0{,}0001$) et explique entre 14 et 29 % de la variance. L'augmentation de la durée estimée manuellement diminue le taux de non-détection. À gauche, les consonnes sonantes génèrent davantage de non-détections que les consonnes voisées qui à leur tour en génèrent davantage que les sourdes. À droite, les consonnes fricatives et liquides donnent lieu à davantage de non-détections que les consonnes nasales ; les consonnes sonantes à davantage de non-détections que les consonnes voisées ou sourdes.

Finalement en ce qui concerne l'alignement IRISA triphones, le modèle prédit les données de manière significative (Chi2 Omnibus : $\chi^2(9) = 432{,}3$, $p < 0{,}0001$), il explique entre 13 et 32 % de la variance. La durée est ici également le facteur le plus important : son augmentation diminue le nombre de non-détections. Concernant la sonorité de la consonne, le profil est similaire à gauche et à droite : les consonnes sonantes donnent lieu à davantage d'erreurs que les consonnes voisées, qui à leur tour génèrent davantage d'erreurs que les consonnes sourdes. Par

ailleurs, une consonne droite nasale génère moins de non-détections qu'une consonne droite liquide ou fricative. Le tableau 3 ci-dessous résume ces données.

Les trois systèmes sont donc fortement similaires en ce qui concerne les variables qui influent sur la détection des schwas présents. La durée est le facteur le plus influent et les traits analysés (mode d'articulation et sonorité) montrent toujours une configuration similaire lorsque leur rôle est significatif, qu'ils appartiennent à la consonne de gauche ou de droite. La sonorité des consonnes a le même impact dans les trois systèmes, l'alignement du LIA diffère quant à l'influence du mode d'articulation. Dans les alignements de l'IRISA, ce dernier joue un rôle dans le contexte droit alors que dans l'alignement du LIA, seule la consonne de gauche est concernée.

Système d'alignement	Consonne gauche	Consonne droite	Durée
LIA	Sonante > voisée > sourde Liquide, fricative > nasale	Sonante > voisée > sourde	–
IRISA monophones	Sonante > voisée > sourde	Sonante > voisée, sourde Fricative, liquide > nasale	–
IRISA triphones	Sonante > voisée > sourde	Sonante > voisée > sourde Fricative, liquide > nasale	–

Tableau 3. *Résumé des variables ayant une influence sur le taux de non-détection pour chacun des systèmes. Les contextes à gauche du signe « > » génèrent davantage d'erreurs que les contextes à sa droite. Pour la durée un « – » indique qu'une augmentation de la durée entraîne une diminution du taux de non-détection*

5.2. *Estimation de la durée du schwa*

5.2.1. *Durées moyennes et écarts*

Outre la détection de la voyelle, il est également intéressant de connaître le comportement d'un système d'alignement quant au placement des frontières des phones. Un décalage important de ces dernières peut avoir des incidences sur les mesures de durée. Les analyses ci-dessous ont été effectuées sur les schwas détectés à la fois par l'alignement automatique et l'alignement manuel de référence.

Le système du LIA attribue au schwa une durée plus importante que celle de l'alignement manuel de référence (55 ms *vs* 52 ms (s = 17)). La différence de durée, évaluée par un test-t apparié, est significative ($t(2\,607) = 10{,}47$, $p < 0{,}0001$). Le système IRISA triphones attribue lui aussi une durée plus importante au schwa que l'alignement manuel (61 ms *vs* 52 ms (s = 18), $t(2\,808) = 26{,}2$, $p < 0{,}0001$). Le système IRISA monophones se distingue en attribuant au schwa une durée inférieure à celle de l'alignement manuel (49 ms *vs* 52 ms (s = 18, $t(2\,744) = 12{,}9$,

$p < 0{,}0001$). Un coefficient de corrélation a également été calculé afin d'évaluer la force de la relation entre l'alignement manuel de référence et chacun des systèmes en ce qui concerne la durée du schwa. Ce coefficient est de 0,65 ($p < 0{,}0001$) pour le système du LIA, de 0,59 ($p < 0{,}0001$) pour IRISA triphones et de 0,72 ($p < 0{,}0001$) pour IRISA monophones. Il est par ailleurs intéressant de relever que malgré l'absence de contraintes concernant la limite temporelle minimale dans le système du LIA, aucune voyelle ne s'est vu attribuer une durée inférieure à 30 ms. Le tableau 4 ci-dessous présente les durées moyennes (et écarts types) des schwas détectés par chacun des systèmes, les pourcentages d'occurrences qui diffèrent par plus de 10 et 20 ms de l'alignement manuel de référence, et les taux d'écarts positifs et négatifs.

Système d'alignement	x(s)	% diff. > 10 ms	% diff. > 20 ms	Écarts positifs	Écarts négatifs
LIA (n = 2 608)	55 (17)	34,4 %	9,8 %	21 9 % > 10 6,2 % > 20	12,5 % > 10 3,6 % > 20
IRISA mono. (n = 2 745)	49 (19)	32,8 %	9,6 %	8,5 % > 10 1,7 % > 20	24 3 % > 10 7,9 % > 20
IRISA tri. (n= 2 809)	61 (21)	44,2 %	17,9 %	37,7 % > 10 15,9 % > 20	6,5 % > 10 2 % > 20

Tableau 4. *Durées moyennes (et écarts types), pourcentages d'écarts supérieurs à 10 et à 20 ms et pourcentages d'écarts positifs et négatifs pour chacun des alignements automatiques*

5.2.2. *Facteurs influençant les écarts de durée*

Nous avons souhaité évaluer dans quelle mesure ces écarts de durée étaient influencés par la nature des consonnes entourant le schwa. Pour ce faire, les écarts ont été classés en trois catégories : durée similaire (à +/– 10 ms) à la durée de référence (= catégorie « correct »), surestimation (> 10 ms) et sous-estimation (< 10 ms). La figure 1 ci-dessous présente le taux d'occurrences dans chaque catégorie pour les trois systèmes.

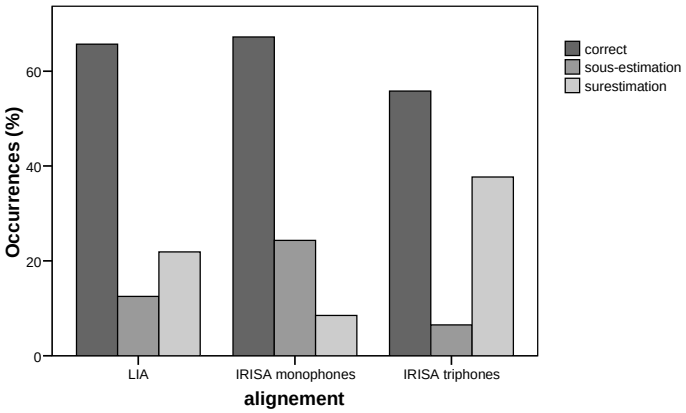

Figure 1. *Pourcentage de surestimations, sous-estimations et estimations correctes de la durée pour les trois systèmes d'alignement évalués*

Comme on peut s'y attendre au regard des résultats présentés dans la section 5.2.1, les alignements des systèmes du LIA et de l'IRISA triphones ont un profil similaire avec un faible nombre de sous-estimations par rapport aux surestimations. L'alignement de l'IRISA monophones se démarque, quant à lui, par un nombre très important de sous-estimations.

Des Chi2 ont été conduits afin d'évaluer la relation entre ces catégories d'estimation et l'entourage consonantique (sonorité et mode d'articulation). Le tableau 6, en annexe, présente les valeurs des Chi2 (et valeurs de p associées) pour chacune de ces analyses. Les quatre types de relations (catégorie/sonorité consonne de droite, catégorie/sonorité consonne de gauche, catégorie/mode consonne de droite, catégorie/mode consonne de gauche) sont toutes significatives pour chacun des alignements considérés. Les principales tendances sont les suivantes. Pour le système du LIA, les sous-estimations sont fréquentes après une consonne sonante et en particulier une liquide. Les surestimations sont fréquentes après une consonne sourde et/ou fricative ou avant une consonne voisée. En revanche, peu de surestimations apparaissent avant une consonne sourde et/ou fricative.

Dans l'alignement du système de l'IRISA monophones, les surestimations sont fréquentes après une consonne sourde et/ou fricative ou avant une consonne sonante, en particulier si elle est liquide. Les surestimations sont, en revanche, peu fréquentes avant une consonne sourde et/ou fricative. Les sous-estimations sont favorisées après une consonne gauche sonante.

En ce qui concerne finalement l'alignement effectué par le système de l'IRISA triphones, ici également les surestimations sont fréquentes après une consonne sourde et/ou fricative, les sous-estimations après une consonne sonante, en particulier liquide. Les sous-estimations sont par ailleurs nombreuses avant une consonne fricative et/ou voisée, elles sont peu nombreuses avant une consonne sonante. Le tableau 5, ci-dessous, résume les propriétés des consonnes ayant une influence sur les taux de surestimation et de sous-estimation. Un « + » signifie que les surestimations sont favorisées par ce contexte, un « – » que les sous-estimations sont favorisées par ce contexte.

	Sonorité gauche			Sonorité droite			Mode gauche				Mode droit			
	Sou.	Son.	V.	Sou.	Son.	V.	F	O	N	L	F	O	N	L
LIA	+	–				+	+			–				
IRISA mono.	+	–			+		+							+
IRISA tri.	+	–				–	+			–	–			

Tableau 5. *Résumé des propriétés consonantiques ayant une influence sur la catégorie d'estimation de la durée (Sou. = sourde, Son. = sonante, V = voisée, F = fricative, O = occlusive, N = nasale, L = liquide)*

En résumé, nous constatons que les trois systèmes sont globalement similaires en ce qui concerne l'influence du type de consonne précédant le schwa sur les taux et types d'erreurs. En revanche, les erreurs des trois systèmes diffèrent dans leurs relations au type de consonne se trouvant après la voyelle. Il s'agit de relever que les analyses statistiques effectuées ici envisagent l'impact de chaque variable séparément, sans tenir compte de l'influence des autres variables.

5.3. *Placement des frontières*

5.3.1. *Début de la voyelle*

Finalement, une analyse du placement des frontières a été entreprise afin de déterminer si les frontières sont décalées de manière égale en début et en fin de voyelle pour chaque système et de comparer ces derniers. La figure 2 ci-dessous présente, pour chaque système d'alignement, le pourcentage des écarts à gauche et à droite, supérieurs à 10 ms et à 20 ms, au regard de l'alignement manuel de référence.

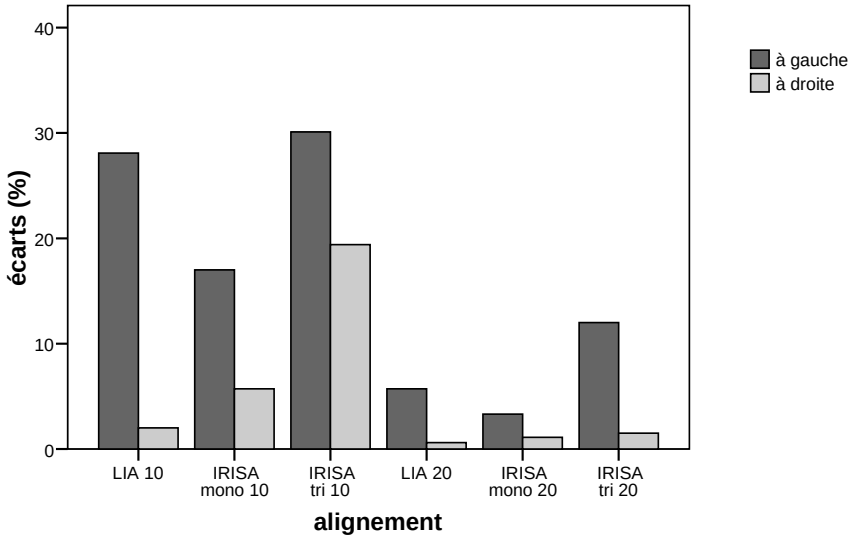

Figure 2. *Pourcentage des écarts à gauche et à droite supérieurs à 10 et à 20 ms en début de voyelle pour chacun des alignements*

Le pourcentage global d'écarts en début de voyelle est élevé pour l'alignement IRISA triphones : 50 % d'écarts supérieurs à 10 ms et 14 % d'écarts supérieurs à 20 ms, comparativement aux performances des deux autres systèmes (LIA : 30 % > 10 ms et 6 % > 20 ms, IRISA monophones : 23 % > 10 ms et 4 % > 20 ms). Pour chacun des systèmes, les écarts sont plus nombreux vers la gauche que vers la droite, la frontière de la voyelle est donc placée plus précocement. Les écarts à gauche oscillent entre 30 % (IRISA triphones) et 17 % (IRISA monophones) à plus de 10 ms, et entre 12 % (IRISA triphones) et 3 % (IRISA monophones) à plus de 20 ms. Intermédiaire, l'alignement du LIA a des valeurs proches de celles de l'IRISA triphones à plus de 10 ms (28 %), mais inférieures quant aux écarts supérieurs à 20 ms (6 %). Les écarts vers la droite sont particulièrement élevés pour l'alignement de l'IRISA triphones dans un intervalle de 10 ms (19 % contre 6 % pour IRISA monophones et 2 % pour LIA). Peu d'écarts à droite à plus de 20 ms sont observés, quel que soit le système considéré (LIA : 0,6 %, IRISA monophones : 1 %, IRISA triphones : 1,5 %).

5.3.2. *Fin de la voyelle*

La figure 3 ci-dessous présente les pourcentages de frontières situées dans un intervalle supérieur à 10 et à 20 ms, à gauche et à droite en fin de voyelle.

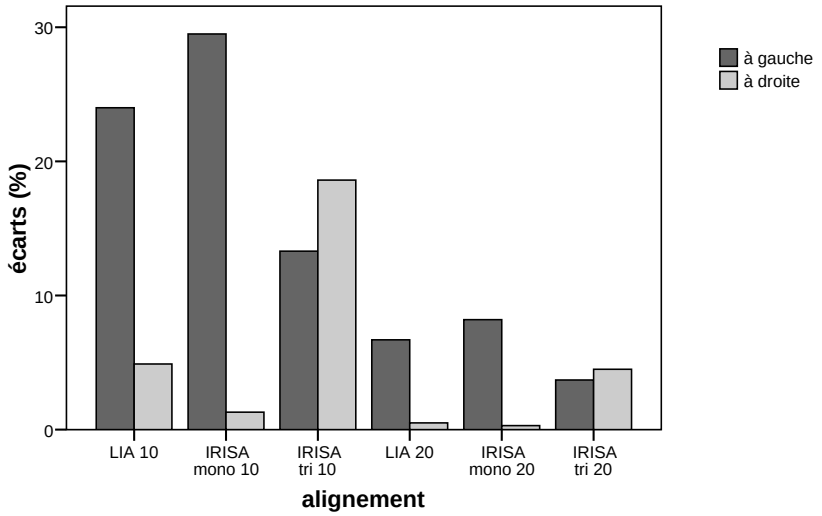

Figure 3. *Pourcentage des écarts à gauche et à droite supérieurs à 10 et à 20 ms en fin de voyelle pour chacun des alignements*

Les taux d'écarts globaux en fin de voyelle sont plus bas qu'en début de voyelle pour l'alignement de l'IRISA triphones (32 % > 10 ms, 8 % > 20 ms). Dans l'alignement de l'IRISA monophones, la configuration inverse est observée, les écarts étant plus nombreux en fin de voyelle (31 % > 10ms, 9 % > 20 ms). En ce qui concerne l'alignement du LIA, le pourcentage d'écarts en fin de voyelle est similaire à celui observé en début de voyelle (30 % > 10 ms, 6 %> 20 ms).

L'alignement du LIA et celui de l'IRISA monophones ont un profil similaire, avec davantage d'écarts vers la gauche que vers la droite. L'alignement « triphones » présente lui davantage d'écarts vers la droite que vers la gauche. Les taux d'écarts vers la gauche varient entre 30 % à plus de 10 ms pour IRISA monophones et 13 % pour IRISA triphones. Ces taux sont respectivement de 8 % et 4 % pour les écarts supérieurs à 20 ms. L'alignement du LIA présente un taux intermédiaire avec 24 % à plus de 10 ms et 7 % à plus de 20 ms. En ce qui concerne les écarts vers la droite, ils sont élevés pour l'alignement IRISA triphones (19 % > 10 ms et 5 % > 20 ms) comparativement aux taux observés pour l'alignement du LIA (5 % et 0,5 % respectivement) et pour l'alignement de l'IRISA monophones (1 % et 0,3 %).

La plus grande durée attribuée aux voyelles par IRISA triphones s'explique donc par un décalage des frontières de début à gauche et de fin à droite. Les alignements du LIA et de l'IRISA monophones ont le même profil, mais la quantité relative des écarts à gauche amène pour l'un à une durée surestimée, pour l'autre à une durée sous-estimée.

6. Discussion

Dans cette étude nous avons souhaité fournir un exemple d'évaluation des implications, pour la recherche en linguistique (phonétique et phonologique en particulier), du recours à des corpus segmentés automatiquement en français. Un cas particulier a été étudié : la voyelle schwa. Plusieurs résultats importants se dégagent de cette analyse, ils sont repris et discutés ci-dessous.

6.1. *Détection du schwa*

Le premier résultat rend compte de la capacité des systèmes d'alignement automatique évalués à décider de la présence/absence d'un phone. Nos données montrent que les systèmes considérés ne sont pas égaux sur ce point. L'alignement effectué par le système de l'IRISA triphones est le plus proche de l'alignement manuel de référence, suivi par celui réalisé par son homologue « monophones ». La comparaison de deux systèmes similaires sur tous les points excepté le type de modèles acoustiques, nous permet d'évaluer la contribution de la nature de ces derniers dans l'adéquation de la détection du schwa. L'utilisation de triphones (modèles de phones dépendants du contexte) plutôt que de monophones en reconnaissance des mots est largement répandue étant donné les meilleures performances qu'elle permet d'obtenir. Le lien entre taux de reconnaissance et détection de phones est relativement direct, une bonne capacité à détecter les phones présents et à ne pas en insérer est cruciale pour éviter les erreurs de reconnaissance (voir cependant Kessens et Strik, 2004). Il paraît de ce fait peu surprenant que le système utilisant des triphones obtienne ici de meilleures performances. Le taux comparativement faible de détection correcte obtenu par le système du LIA, utilisant lui aussi des triphones en interne de mots, est difficile à expliquer, d'autant plus que la performance du système en reconnaissance est relativement bonne (Galliano *et al*., 2005). Il s'agirait ici d'étendre l'analyse à d'autres systèmes/paramétrages, afin de déterminer si la différence de performance peut être expliquée par le nombre de gaussiennes ou la qualité de la phonétisation, paramètres qui diffèrent entre le système du LIA et celui de l'IRISA triphones. Peut-être faut-il préciser ici que, s'agissant du schwa, les erreurs de détection à l'intérieur des mots n'ont pas autant d'impact sur le taux de reconnaissance correcte des mots que lorsque d'autres phones sont impliqués.

De manière générale cependant, les trois systèmes témoignent de performances, en termes de détection de phones, tout à fait acceptables au regard de ce qui est présenté dans la littérature. (Van Bael *et al*., 2007) rapportent, par exemple, des pourcentages de désaccord impliquant des insertions et effacements de phones de 6 à 8 % (tous phonèmes confondus). Ces valeurs sont proches des 7 à 10 % obtenus ici. Les coefficients de Kappa obtenus par (Kessens et Strik, 2004) pour le traitement de l'effacement et de l'insertion du schwa en néerlandais, sont eux

beaucoup plus bas que ceux obtenus par les trois systèmes évalués ici. Relevons finalement que le degré d'accord manuel/automatique que nous obtenons n'est jamais aussi bon que celui observé entre les deux alignements manuels. Si cet effet est probablement renforcé par la démarche suivie (voir point 4.2.4.), cette constatation est récurrente dans la littérature (Wester *et al.*, 2001). Les implications des divergences observées entre alignement manuel et automatique en ce qui concerne l'étude du schwa sont *a priori* difficiles à estimer. Elles vont dépendre, outre de l'objectif de l'analyse, de l'existence ou non de biais systématiques intervenant dans les décisions des systèmes automatiques. Les analyses des facteurs impliqués ont été conduites dans le but de déterminer l'existence de tels biais.

La seconde série d'analyses visait à évaluer l'influence potentielle de facteurs segmentaux sur les taux d'erreurs de détection. Nous observons une configuration sensiblement différente selon que l'analyse porte sur les schwas présents ou les schwas absents dans l'alignement manuel de référence. Lorsque le schwa est absent, l'entourage consonantique n'influence pas les trois systèmes de façon égale. Cependant, la faible capacité des facteurs considérés à rendre compte des données suggère que d'autres facteurs sont également impliqués et limite la portée des observations effectuées. En ce qui concerne les schwas présents, les trois systèmes montrent des configurations fortement similaires. La durée du schwa estimée manuellement est un fort prédicteur du taux d'erreurs : plus elle augmente, plus les non-détections diminuent. L'influence de la sonorité des consonnes et du mode d'articulation est également similaire dans les trois alignements. Les consonnes sonantes génèrent davantage de non-détections que les consonnes voisées qui, à leur tour, génèrent davantage de non-détections que les consonnes sourdes. Ce résultat n'est pas surprenant au regard des similarités/différences acoustiques qui peuvent exister entre ces consonnes et le schwa. Les consonnes sonantes présentant des formants, lorsqu'elles suivent ou précèdent une voyelle, il peut être difficile de distinguer un changement net au moment où commence/termine cette dernière. En revanche, la distinction entre une consonne sourde et une voyelle est beaucoup plus aisée, les caractéristiques acoustiques de ces deux segments étant très différentes. Bien que cela nécessite d'être entériné par l'évaluation d'un nombre plus important de systèmes d'alignement, nous pouvons faire l'hypothèse que les régularités dégagées se retrouveront dans un grand nombre d'entre eux.

Les taux d'erreurs de détection du schwa, insertions et non-détections confondues, oscillent entre 12 et 19 % selon le système d'alignement considéré. Ce taux n'est pas négligeable, et le fait que l'adéquation de la détection soit influencée par l'environnement consonantique ne doit pas être occulté suivant le type d'étude linguistique envisagé. Ainsi, une analyse des contextes qui favorisent ou non la chute du schwa, thème récurrent dans la littérature linguistique, sera susceptible de rendre compte autant des caractéristiques des systèmes automatiques que de facteurs linguistiques si elle se fonde sur un alignement non vérifié manuellement. Une analyse des caractéristiques temporelles du schwa risque également de ne pas

refléter la réalité acoustique, puisque les schwas les plus courts ont tendance à ne pas être détectés par les différents systèmes.

6.2. *Critères temporels*

Si une détection adéquate est capitale en reconnaissance des mots, lorsque l'on s'intéresse aux réalisations acoustiques des phonèmes, d'autres critères sont également à considérer. L'emplacement des frontières en particulier peut s'avérer crucial suivant les analyses envisagées. L'une des conséquences d'une détermination inadéquate de l'emplacement des frontières du phone peut être une mauvaise évaluation de la durée de ce dernier. Nous avons observé des différences parfois importantes entre l'estimation manuelle et l'estimation automatique comme en témoigne l'exemple présenté dans la figure 4 ci-dessous. Si, en termes de détection, le système du LIA fait office de mauvais élève, en ce qui concerne l'estimation de la durée du schwa, le système IRISA triphones obtient la moins bonne performance. Le système IRISA monophones sort gagnant de cette comparaison. Une explication peut être avancée ici, impliquant à nouveau le type de modèles de phones utilisé par les différents systèmes. Lorsque des modèles de phones dépendants du contexte sont utilisés, les frontières entre les phones sont plus « floues » : il est difficile de savoir quelle partie du phone va être considérée comme telle ou comme partie du contexte lors de l'apprentissage. On peut dès lors s'attendre à davantage d'erreurs d'estimation de la durée. La moins bonne performance des systèmes utilisant des triphones a d'ailleurs déjà été rapportée dans la littérature (Lanchantin *et al.*, 2008). La performance intermédiaire du système du LIA pourrait se trouver expliquée par l'influence des modèles de phones indépendants du contexte pour les mots ayant un schwa en seconde position. Il est également à relever que le système utilisant des modèles de phones indépendants du contexte attribue au schwa une durée plus courte que l'alignement manuel alors que la durée attribuée par les systèmes pourvus de modèles de phones dépendants du contexte est plus grande que celle de l'alignement manuel. La sur/sous-estimation systématique des durées est probablement liée, outre à la qualité des modèles de phones, à la nature des phonèmes. Si le schwa amène à une surestimation de la durée, d'autres phonèmes seront systématiquement sous-estimés et inversement, puisque la durée totale du mot doit respecter celle de la variante de prononciation et ne peut être étendue au-delà de ses limites.

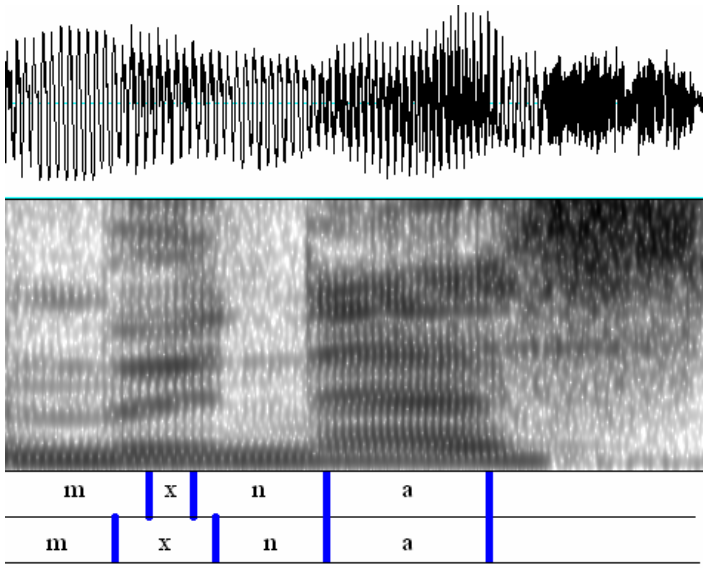

Figure 4. *Exemple de différence dans le placement des frontières du schwa (« x ») entre l'alignement automatique effectué par le système « triphones » de l'IRISA (en haut) et l'alignement manuel (en bas) pour une occurrence du mot « menace »*

Les trois systèmes sont fortement similaires en ce qui concerne l'influence du type de consonne précédant le schwa sur les taux et types d'erreurs. Une consonne sourde et/ou fricative génère de nombreuses surestimations de la durée, alors que la présence de consonnes sonantes tend à augmenter le nombre de sous-estimations. On remarque que les consonnes favorables aux sous-estimations sont les mêmes que celles favorisant les non-détections. Ici à nouveau les caractéristiques acoustiques du schwa et leurs similarités/différences avec celles des consonnes peuvent être invoquées. Les influences ne vont cependant pas toujours dans le même sens pour les trois systèmes. Non seulement les facteurs impliqués diffèrent, mais lorsqu'ils sont impliqués dans plusieurs systèmes, leurs influences peuvent être contradictoires. Par exemple, une consonne voisée à droite n'influence pas l'estimation de la durée pour le système de l'IRISA monophones ; elle favorise les surestimations pour l'alignement du LIA et les sous-estimations pour l'alignement IRISA triphones. Il s'agit cependant de préciser ici que si les valeurs relatives des Chi2 nous donnent un premier aperçu de l'importance des facteurs envisagés, la colinéarité potentielle de ces facteurs n'est pas prise en considération. Étant donné l'absence d'équilibre dans nos contextes, ces conclusions doivent de ce fait être pondérées : il est possible que le rôle de certains facteurs soit moins important lorsque celui des autres est également considéré. Ces résultats vont néanmoins dans

le sens de ce qui est observé dans la littérature. (Wesenick et Kipp, 1996), bien que ne s'intéressant qu'aux frontières entre deux consonnes, relèvent également que lorsque des consonnes nasales ou liquides sont impliquées, le système d'alignement automatique rencontre davantage de difficultés. Nous constatons comme cet auteur que les segmentations difficiles pour le transcripteur manuel le sont également pour la machine. Dans l'étude de (Nguyen et Espesser, 2004), les mauvaises performances sont liées aux contextes droits /ʀ, j, z/. Ce profil se rapproche de ce qui est observé ici pour le système IRISA monophones.

En ce qui concerne finalement le placement des frontières en début et fin de voyelle, rappelons que dans l'étude de (Nguyen et Espesser, 2004), les écarts étaient plus importants en fin de voyelle qu'en début. Nos résultats montrent que cela ne peut être généralisé à tous les systèmes. Si l'alignement « monophones » de l'IRISA suit ce profil[8], on n'observe pas de différence marquée pour l'alignement du LIA, et un profil inverse pour l'alignement IRISA triphones. Par ailleurs, les trois systèmes d'alignement considérés tendent à placer la frontière de début de voyelle plus tôt que le transcripteur manuel. La frontière de fin de voyelle est placée plus souvent à droite par le système du LIA, et à gauche par les deux autres systèmes. Signalons encore qu'au regard des performances relatées dans la littérature, les deux systèmes de l'IRISA obtiennent une performance souvent meilleure en ce qui concerne l'emplacement des frontières, alors que celle du LIA est inférieure à ce qui est souvent rapporté.

Le choix du système va donc ici également influencer la pertinence de la segmentation et par là son adéquation à des fins d'analyse linguistique. L'objectif de l'étude doit à nouveau être considéré. S'il s'agit d'étudier les caractéristiques temporelles de la voyelle, les données issues d'un alignement automatique devront être considérées avec prudence. Nous avons vu en effet que les durées estimées automatiquement diffèrent parfois fortement des durées segmentées manuellement et qu'elles sont influencées par les consonnes suivantes et précédentes. L'étude de l'influence du contexte segmental sur la durée des voyelles en particulier risque d'être fortement biaisée si elle s'appuie sur un alignement non vérifié manuellement. Par ailleurs, il s'agit de garder à l'esprit les limitations imposées par le système à la durée d'un segment, que cette limite soit ou non imposée par les modèles de phones. Dans les alignements automatiques évalués ici, la voyelle ne se voit jamais attribuer une durée inférieure à 30 ms, or, la durée minimale attribuée au schwa par l'alignement manuel est de 8 ms. En ce qui concerne l'impact des divergences temporelles entre l'alignement manuel et l'alignement automatique sur des analyses formantiques, une étude a été entreprise par (Adda-Decker *et al*., ce volume). Si les imprécisions d'alignement sont généralement corrigées dans les

8. Il serait intéressant ici de connaître le type de modèles de phones utilisés dans cette étude. L'alignement du système IRISA monophones semble se rapprocher davantage des résultats qui y sont mentionnés que l'alignement des deux autres systèmes, basés sur des modèles de phones dépendants du contexte.

analyses de par la grande quantité de données, les auteurs suggèrent de prendre certaines précautions méthodologiques. Une analyse acoustique des voyelles dans la partie médiane (du premier au dernier tiers) restera assez peu sensible aux imprécisions de la segmentation, mais il n'en sera pas de même pour une analyse visant à analyser des voyelles plus courtes ou des parties spécifiques des voyelles (transition consonne-voyelle ou voyelle-consonne, par exemple).

7. Conclusion

Cette étude s'est attachée à montrer les différences entre les décisions de systèmes automatiques et de transcripteurs manuels. Les biais éventuels ont été traqués, et pour certains d'entre eux démasqués. Le risque de conclure à des généralités linguistiques qui sont en fait davantage liées à l'outil utilisé existe, il a été souligné. Doit-on pour autant renoncer à ce type d'outils dans le contexte d'analyses linguistiques fines ? Si la faillibilité des systèmes d'alignement automatique est soulignée à plusieurs reprises dans la littérature, leurs avantages sont également fréquemment rapportés. Outre la quantité de matériel qu'ils permettent de traiter et leur faible coût, déjà mentionnés, l'absence de subjectivité et l'uniformité de leurs décisions sont également mises en avant (Cucchiarini et Strik, 2003). Afin qu'ils demeurent un outil privilégié cependant, permettant de conduire à des études de qualité, leur emploi doit se faire en connaissance de cause. Il paraît notamment nécessaire que les phonéticiens se renseignent sur les caractéristiques et performances du système qu'ils envisagent d'utiliser et qu'ils évaluent son adéquation pour la tâche envisagée. Par ailleurs, une certaine quantité de données devrait être soumise à une vérification manuelle.

Quelques généralisations peuvent être dégagées de nos données, susceptibles de guider les linguistes dans le choix et l'évaluation de systèmes d'alignement automatique. L'impact de l'entourage consonantique et de la durée sur la détection et le placement des frontières souligne la nécessité de tenir compte de ces aspects mais permet également de poser des hypothèses sur la direction des biais éventuels. Afin d'entériner ces résultats, davantage d'études sont cependant encore nécessaires. Il s'agirait notamment d'étudier d'autres contextes et d'apporter une quantité plus importante de détails concernant le rôle particulier des différentes consonnes à l'intérieur des grandes classes dégagées. Par ailleurs, des analyses statistiques supplémentaires permettraient de rendre compte des éventuelles interactions entre les facteurs étudiés. Il a également été montré que certains systèmes/paramétrages étaient plus adaptés à certaines tâches. Ainsi, si un taux optimal de détection est souhaité, il est préférable d'opter, toutes choses étant par ailleurs égales, pour les modèles de phones dépendants du contexte. Si en revanche une plus grande précision temporelle est nécessaire, les systèmes ayant recours à des modèles de phones indépendants du contexte semblent plus appropriés.

Par ailleurs, des liens entre les alignements et les caractéristiques des systèmes qui les ont engendrés ont été évoqués. L'importance notamment de la nature des modèles de phones a été soulignée à plusieurs reprises. Ces liens demandent à être étoffés et approfondis, le rôle d'autres paramètres (nombre de gaussiennes, etc.) doit être évalué. Ces liens suggèrent qu'il est possible d'envisager le développement d'outils spécifiques pour répondre aux impératifs des analyses linguistiques. La littérature regorge d'articles témoignant de tentatives d'améliorer les taux de reconnaissance. Un bon système de reconnaissance n'est cependant pas forcément un bon outil d'alignement, et rien ne permet de penser que l'amélioration des taux de reconnaissance de mots ira de pair avec l'apparition d'outils d'alignement plus efficaces (Kessens et Strik, 2004). Des systèmes d'alignement automatique doivent être optimisés pour cette tâche (Cucchiarini et Strik, 2003). Notons, cependant, que l'optimisation d'aligneurs pour les besoins du TAP ne garantit pas l'émergence d'outils adaptés aux linguistes. Comme le soulignent (van Bael *et al.*, 2007), un alignement optimal pour le TAP n'est pas forcément celui qui ressemble le plus à un alignement manuel. L'alignement effectué dans le cadre de la synthèse par sélection d'unités pourrait s'avérer plus pertinent (Golipour et O'Shaughnessy, 2007 ; Kuo *et al.,* 2007) que les alignements utilisés en reconnaissance des mots. L'idéal serait cependant que l'optimisation d'un système d'alignement à des fins linguistiques soit entreprise dans une collaboration rapprochée entre chercheurs des deux disciplines.

Si les avantages d'une telle entreprise pour les linguistes sont évidents, la recherche en TAP devrait également en sortir gagnante. La quantité de savoir phonétique utilisée actuellement en TAP dans les systèmes automatiques est plutôt réduite. Si les performances sont satisfaisantes pour de la parole lue ou de la parole non lue soignée (environ 10 % d'erreurs de reconnaissance des mots selon la dernière campagne d'évaluation ESTER), le traitement automatique de la parole spontanée n'en est qu'à ses prémices. Élisions, réductions et dysfluences sont de nombreuses sources de variation dans la parole spontanée auxquelles se heurtent les systèmes. Plusieurs auteurs suggèrent que les améliorations futures en TAP passeront par la prise en compte de facteurs linguistiques (Strik, 2005 ; Pols, 1999) en particulier en ce qui concerne la variation phonologique. Ils soulignent cependant que le transfert du savoir phonétique actuel est rendu difficile par le fait que ce dernier est fondé sur de faibles quantités de parole généralement produite en laboratoire et non généralisable à la parole continue. Or, si les linguistes disposent d'outils performants permettant un alignement précis de grands corpus de parole continue, ces obstacles vont s'élimer. Des connaissances phonétiques adaptées aux besoins du TAP et utilisables par ce dernier pourront alors voir le jour.

Que ce soit pour l'élaboration d'outils pour des besoins spécifiquement linguistiques ou dans l'optique d'une meilleure appréhension par les linguistes, des caractéristiques et implications des alignements automatiques existants, il est nécessaire de poursuivre les investigations ébauchées dans le cadre de ce travail. Les évaluations ont concerné ici uniquement la capacité des systèmes à rendre

compte du schwa. Il convient de ne pas oublier que le schwa est une des voyelles, sinon la voyelle, qui pose le plus de problèmes de détection et d'alignement. Les résultats présentés ici seraient probablement meilleurs pour les autres voyelles, cela reste cependant à entériner dans une étude ultérieure. De même, les régularités observées ainsi que le rôle des différents facteurs avancés demandent à être évalués pour les autres segments du français. Finalement, la présente étude s'est limitée à considérer l'influence de facteurs locaux sur l'alignement. Il s'agirait de poursuivre l'analyse en incluant des facteurs de plus haut niveau, notamment prosodiques.

	LIA (n = 2 608)	Monophones (n = 2 745)	Triphones (n = 2 809)
Sonorité droite	χ^2 (4) = 11,2, p < 0,05	χ^2(4) = 30,52, p < 0,0001	χ^2 (4) = 39,6, p < 0,0001
Sonorité gauche	χ^2 (4) = 36,24, p < 0,0001	χ^2(4) = 210,85, p < 0,0001	χ^2 (4) =186,9, p < 0,0001
Mode droite	χ^2 (6) = 24,65, p < 0,0001	χ^2(6) = 61,02, p < 0,0001	χ^2 (6) = 53,4, p < 0,0001
Mode gauche	χ^2 (6) = 62,2, p < 0,0001	χ^2(6) = 263,4, p < 0,0001	χ^2 (6) = 151,8, p < 0,0001

Tableau 6. *Valeurs des Chi2 pour chaque système, test de la relation entre catégorie d'estimation de la durée et contextes gauche et droit*

Remerciements

Les auteurs souhaitent remercier Julien Chanal pour ses conseils concernant les procédures statistiques, Sandra Schwab ainsi que deux relecteurs anonymes pour leurs commentaires et suggestions lors de la lecture d'une version antérieure de l'article.

8. Bibliographie

Adda-Decker, M., « Problèmes posés par le schwa en reconnaissance et en alignement automatiques de la parole », *Actes des 5e journées d'études linguistiques*, Nantes, France, juin 2007, p. 211-216.

Adda-Decker, M., Lamel, L., « Systèmes d'alignement automatique et études de variantes de prononciation », *Actes des 23e journées d'études sur la parole*, Aussois, France, juin 2000, p. 189-192.

Adda-Decker, M., Gendrot, C., Nguyen, N., « Apport du traitement automatique à l'étude des voyelles », *Revue T.A.L*, vol. 49, n° 3, 2008.

Auran, C., Bouzon, C., « Phonotactique prédictive et alignement automatique : application au corpus MARSEC et perspectives », *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence*, vol. 22, 2003, p. 33-63.

Béchet, F., « LIA_PHON : un système complet de phonétisation de textes », *Revue T.A.L.*, vol. 42, n° 1, 2001, p. 47-67.

Binnenpoorte, D., Cucchiarini, C., Strik, H., Boves, L., « Improving automatic phonetic transcription of spontaneous speech through variant-based pronunciation variation modelling », *Proceedings of LREC*, Lisbonne, Portugal, mai 2004, p. 681-684.

Boersma, P., Weenink, D., Praat : doing phonetics by computer. (Version 4.6.15) http://www.praat.org/, 2007.

Boula de Mareüil, P., Yvon, F., d'Alessandro, C., Auberge, V., Vaissière, J., Amelot, A., « A French phonetic lexicon with variants for speech and language processing », *Proceedings of LREC*, Athènes, Grèce, juin 2000, p. 273-276.

Boula de Mareüil, P., Adda-Decker, M., « Studying pronunciation variants in French by using alignment techniques », *Proceedings of Interspeech*, Denver, USA, sept. 2002, p. 2273-2276.

Brugnara, F., Falavigna D., Omologo, M., « Automatic segmentation and labeling of speech based on Hidden Markov Models », *Speech Communication*, vol. 12, n° 4, 1993, p. 357-370.

Bürki, A., Fougeron, C., Gendrot, C., Frauenfelder, U., « De l'ambiguïté de la chute du schwa en français », *Actes des 5e journées d'études linguistiques*, juin 2007, Nantes, France, p. 83-88.

Bürki, A., Fougeron, C., Gendrot, C., « On the categorical nature of the process involved in schwa elision in French », *Procceedings of Interspeech*, Anvers, Belgique, sept. 2007, p. 1026-1029.

Cohen, J., « Weighted kappa : nominal scale agreement with provision for scaled disagreement or partial credit », *Psychological Bulletin*, vol. 70, 1968, p. 213-220.

Content, A., Mousty, P., Radeau, M., « Brulex, une base de données lexicales informatisée pour le français écrit et parlé », *L'Année Psychologique*, vol. 90, 1990, p. 551-566.

Côté, M., Morrison, G., « The nature of the schwa/zero alternation in French clitics : experimental and non-experimental evidence », *Journal of French and Language Studies*, vol. 17, 2007, p. 159-186.

Cucchiarini, C., Strik, H., « Automatic phonetic transcription : an overview », *Proceedings of ICPHS*, Barcelone, Espagne, Août 2003, p. 347-350.

Fougeron, C., Gendrot, C., Bürki, A., « On the phonetic identiy of French schwa, compared to /ø/ and /œ/ », *Actes des 5e journées d'études linguistiques,* juin 2007, Nantes, France, p. 83-88.

Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.-F., Gravier, G., « ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News », *Proceedings of Interspeech*, Lisbonne, Portugal, sept. 2005, p. 1149-1152.

Gendrot, C., Adda-Decker, M., « Impact of duration on F1/F2 formant values of oral vowels : an automatic analysis of large broadcast news corpora in French and German », *Proceedings of Interspeech*, Lisbonne, Portugal, sept. 2005, p. 2453-2456.

Golipour, L., O'Shaughnessy, D., « A new approach for phoneme segmentation of speech signals », *Procceedings of Interspeech*, Anvers, Belgique, sept. 2007, p. 1933-1936.

Kessens, J., Strik, H., « Lower WERs do not guarantee better transcriptions », *Proceedings of Interspeech*, Aalborg, Danemark, sept. 2001, p. 1721-1724.

Kessens, J., Strik, H., « On automatic phonetic transcription quality : lower word error rates do not guarantee better transcriptions », *Computer Speech and Language*, vol. 18, 2004, p. 123-141.

Kuo, J., Lo, H., Wang, H., « Improved HMM/SVM methods for automatic phoneme segmentation », *Procceedings of Interspeech*, Anvers, Belgique, sept. 2007, p. 2057-2060.

Kuperman, V., Pluymaekers, M., Ernestus, M., Baayen, H., « Morphological predictability and acoustic duration of interfixes in Dutch compounds », *Journal of the Acoustic Society of America*, vol. 121, n° 4, 2007, p. 2261-2271.

Landis, J., Koch, G., « The measurement of observer agreement for categorical data », *Biometrics,* vol. 33, 1967, p. 159-174.

Lanchantin, P., Morris, A., Rodet, X., Veaux, C., « Automatic phoneme segmentation with relaxed textual constraints », *Proceedings of LREC 08,* Marrakech, Maroc, mai 2008.

Malécot, A., *Introduction à la phonétique française*, The Hague, Mouton, 1977.

New, B., Pallier, C., Ferrand, L., Matos, R., « Une base de données lexicales du français contemporain sur internet : LEXIQUE », *L'Année Psychologique*, vol. 101, 2001, p. 447-462.

Nguyen, N., Espesser, R., « Méthodes et outils pour l'analyse acoustique des systèmes vocaliques », *Bulletin Phonologie du français contemporain*, vol. 3, 2004, p. 77-85.

Pitt, M., Johnson, K., Hume, E., Kiesling, S., Raymond, W., « The Buckeye corpus of conversational speech : labeling conventions and a test of transcriber reliability », *Speech Communication*, vol. 45, 2005, p. 89-95.

Pols, L., « Flexible, robust and efficient human speech processing versus present-day technology », *Proceedings of ICPHS*, San Francisco, USA, août 1999, p. 9-16.

Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A., McDonough, J., Nock, H., Saraclar, M., Wooters, C., Zavaliagkos, G., « Stochastic pronunciation modelling from hand-labelled phonetic corpora », *Speech Communication*, vol. 29, 1999, p. 209-224.

Shriberg, L., Kwiatkowski, J., Hoffmann, K., « A procedure for phonetic transcription by consensus », *Journal of Speech and Hearing Research*, vol. 27, 1984, p.456-465.

Sjölander, K., « An HMM-based system for automatic segmentation and alignment of speech », *Phonum*, vol. 9, 2003, p. 93-96.

Stemler, S., « A comparison of consensus, consistency and measurement approaches to estimating interrater reliability », *Practical Assessment, Research and Evaluation*, vol. 9, n° 4, 2004, Retrieved January 20, 2008 from http://PAREonline.net/getvn.asp?v=9&n=4

Strik, H., « Is phonetic knowledge of any use for speech technology? », In B. Barry et W. van Dommelen (Eds), *The integration of phonetic knowledge in speech technology*, Series : Text, Speech and language technology, vol. 25, Springer, Dordrecht, 2005, p. 167-180.

Strik, H., Elffers, A., Bavcar, D., Cucchiarini, C., « Half a word is enough for listeners, but problematic for ASR », *Proceedings of ITRW on Speech Recognition and Intrinsic Variation,* Toulouse, France, mai 2006, p. 101-106.

Strik, H., Cucchiarini, C., « Modeling pronunciation variation for ASR : a survey of the literature », *Speech Communication*, vol. 29, 1999, p. 225-246.

van Bael, C., van den Heuvel, H., Strik, H., « Validation of phonetic transcriptions in the context of automatic speech recognition », *Language Ressources and Evaluation*, vol. 41, n° 2, 2007, p. 129-146.

van Bael, C., Baayen, H., Strik, H., « Segment deletion in spontaneous speech : a corpus study using mixed effects models with crossed random effects », *Proceedings of Interspeech*, Anvers, Belgique, août 2007, p. 2741-2744.

Wesenick, M., Kipp, A., « Estimating the quality of phonetic transcriptions and segmentations of speech signals », *Proceedings of ICSLP*, Philadelphia, USA, oct. 1996, p. 129-132.

Wester, M., Kessens, J., Cucchiarini, C., Strik, H., « Obtaining phonetic transcriptions : a comparison between expert listeners and a continuous speech recognizer », *Language and Speech*, vol.44, n° 3, 2001, p. 377-403.