Fouille de données audio pour la classification automatique de mots homophones

Rena NEMOTO*, Martine Adda-Decker*
Ioana Vasilescu*

*LIMSI-CNRS B.P. 133 91403 Orsay Cedex France
{nemoto, madda, ioana}@limsi.fr
http://www.limsi.fr

Résumé. Cet article présente une contribution à la modélisation acoustique des mots à partir de grands corpus oraux, faisant appel aux techniques de fouilles de données. En transcription automatique, de nombreuses erreurs concernent des mots fréquents homophones. Deux paires de mots (quasi-)homophones *à*/*a* et *et*/*est* sont sélectionnées dans les corpus, pour lesquels sont définis et examinés 41 descripteurs acoustiques permettant potentiellement de les distinguer. 17 algorithmes de classification, mis à l'épreuve pour la discrimination automatique de ces deux paires de mots, donnent en moyenne 77% de classification correcte sur les 5 meilleurs algorithmes. En réduisant le nombre de descripteurs à 10 (sélectionnés par l'algorithme le plus performant), les résultats de classification restent proches du résultat obtenu avec 41 attributs. Cette comparaison met en évidence le caractère discriminant de certains attributs, qui pourront venir enrichir à la fois la modélisation acoustique et nos connaissances des prononciations de l'oral.

1 Introduction

En transcription automatique de la parole, de grands corpus audio (incluant généralement des centaines d'heures de parole) servent à estimer des modèles acoustiques précis de phonèmes contextuels. Ces modèles de sons élémentaires sont ensuite concaténés pour aboutir à des modèles de mots en s'appuyant sur la connaissance de leur prononciation. Cette connaissance est incomplète à l'heure actuelle et une partie importante de l'information caractérisant les variantes de prononciations se trouve encodée implicitement dans les modèles acoustiques. L'objectif de ce travail est de s'appuyer sur les techniques de fouille de données afin d'extraire des connaissances relatives aux spécificités acoustiques et prosodiques caractérisant les prononciations. Cette approche a déjà pu montrer son intérêt pour la caractérisation des accents étrangers (Vieru-Dimulescu et al., 2007). Nous nous intéresserons ici aux mots considérés comme homophones, i.e. phonémiquement pareils, et qui sont de ce fait sujets à de nombreuses erreurs de confusion lors de la transcription automatique. Partant de ces constats, nous nous sommes interrogés si les mots homophones ne déploieraient pas de particularités acoustiques/prosodiques qui n'ont été prises en compte ni par les paramètres acoustiques classiques (vecteurs de cepstres), ni par les modèles acoustiques (Modèles de Markov Cachés à trois états) et qui permettrait leur discrimination. Nous faisons ainsi l'hypothèse que des informations prosodiques (concernant durée, fréquence fondamentale notée f0, cooccurrence avec des pauses, etc.) puissent contribuer à lever certains types d'homophonie, en particulier s'il s'agit d'homophones issus de classes syntaxiques différentes (hétéro-syntaxiques). Nous avons fait appel aux techniques de fouille de données afin de classer automatiquement ces

mots grâce à un ensemble d'attributs acoustiques/prosodiques spécifiques développés pour ce travail.

Comme déjà évoqué ci-dessus, de nombreuses erreurs de transcription concernent des mots, considérés comme homophones, par exemple une confusion entre un nom au singulier et un nom au pluriel (*table*, *tables*), un verbe au participe passé ou à l'infinitif (*allé*, *aller*) ou des mots-outils homophones (*à*, *a*). Ces derniers, par leur fréquence d'occurrences dans la langue participent de manière significative aux erreurs de transcription automatique.

Dans la section 2, nous allons présenter les corpus utilisés et les analyses menées concernant : la fréquence des mots, la durée, la f0, en nous limitant à deux paires de mots homophones choisies pour cette étude préliminaire, i.e. *à* (préposition) vs. *a* (verbe) et *et* (conjonction) vs. *est* (verbe). Ces mesures visent à étudier les réalisations acoustiques de ces mots homophones hétéro-syntaxiques dans le but d'identifier d'ores et déjà des attributs potentiellement pertinents lors de la classification automatique. Nous allons ensuite essayer de mettre en évidence des traits spécifiques différenciant ces mots fréquents en utilisant la classification automatique et la fouille de données (section 3), avant de conclure et d'ouvrir quelques perspectives (section 4).

2 Analyses acoustiques de mots (quasi-)homophones

Lors de la campagne ESTER (Evaluation des Systèmes de Transcription enrichie d'Emissions Radiophoniques) (Galliano et al., 2005), financée par le programme interministériel français TECHNOLANGUE et organisée par l'AFCP (Association Francophone de la Communication Parlée), la DGA (Délégation Générale pour l'Armement) et ELDA (European Language Resources Distribution Agency), le système de transcription automatique du LIMSI a obtenu environ 11% de taux d'erreur de mots (Galliano et al., 2005). Cependant de nombreuses erreurs portent sur les mots outils, qui sont les mots les plus fréquents et qui sont de plus souvent monosyllabiques comme *et*, *est*, *a*, *à*, *un, que*, *q*ui, *il, y, etc.* (Adda-Decker, 2006). Dans l'étude présentée ici, deux paires de mots *et* (conjonction)/*est* (verbe *être*) et *à* (préposition)/*a* (verbe *avoir*), qui sont parmi les mots les plus fréquents du français et qui sont souvent confondues lors de la transcription automatique, sont choisies pour définir et examiner des descripteurs ou attributs acoustiques permettant potentiellement de distinguer ces paires. Il faut souligner que la paire *et*/*est* n'est pas vraiment homophone au sens phonologique, car le /E/ y correspond à deux degrés d'aperture différents: la réalisation /e/ (e fermé) caractérise le mot *et*, tandis que la prononciation canonique du verbe *est* est /ɛ/ (e ouvert). Cependant, dans la parole spontanée, la réalisation acoustique en tant que [e] (fermé) du verbe est fréquente, ce qui entraîne l'homophonie des deux mots dans ce cas.

2.1 Corpus utilisés

Cette étude part de deux types de corpus en français : l'un est composé de journaux radiodiffusés d'environ 55 heures et appelé **BN** (*Broadcast News*), provenant de différentes stations de radios (France Inter, Radio France International (RFI), France Info et Radio Television du Maroc (RTM)). Le style de ces corpus est (semi-) préparé. L'autre corpus, appelé **PFC** (Phonologie du Français Contemporain) (Durand et al., 2003) contient des enregistrements de variétés de français de régions différentes et en différents styles de parole : parole spontanée (entretiens) et lecture (liste de mots et texte). La partie du corpus PFC utilisé contient environ 32 heures, dont le style « entretien » couvre 20h. Nous avons retenu ce sous-ensemble d'entretiens correspondant à de l'oral spontané.

2.2 Alignement automatique

L'alignement automatique vise à localiser dans le signal acoustique les mots prononcés, de déterminer leur prononciation et de segmenter le flux audio en phones. Pour ce faire le système d'alignement utilise des transcriptions manuelles, un dictionnaire de prononciation et des modèles acoustiques de phones indépendants du contexte. Le système de transcription automatique de parole du LIMSI (Gauvain et al., 2005) est utilisé pour l'alignement des corpus décrits ci-dessus. Le but est de localiser les paires *à*/*a* et *et*/*est* dans le flux de parole. L'alignement permet également de localiser les pauses. Concernant le mot *est,* le dictionnaire de prononciation inclut plusieurs variantes ([ɛ], [e], [ɛt], [et]) permettant au-delà de la prononciation canonique, une réalisation avec le [e] fermé et la réalisation de la liaison. L'alignement automatique permet de mesurer la durée de nos deux paires de mots, ainsi que des mesures de durées contextuelles. La résolution temporelle de la segmentation est limitée à 10*ms* et la durée minimale d'un segment est de 30*ms* pour des raisons techniques.

2.3 Extraction automatique des paramètres acoustiques

Le logiciel PRAAT (Boersma et Weenink, 1999) a été utilisé afin d'extraire un nombre de paramètres acoustiques intervenant par la suite dans la définition d'attributs pour la classification des mots homophones. Ces paramètres concernent le mot même (mot cible) et le contexte immédiat gauche/droite (i.e. la voyelle ou la pause précédant/suivant le mot cible). Nous avons ainsi extrait les trois premiers formants (F1, F2, F3) et la fréquence fondamentale (f0). Les formants (F1, F2 et F3) apportent des informations relatives à la qualité vocalique des segments: globalement le premier formant (F1) correspond à l'articulation sur un axe ouvert/fermé du segment, le second (F2) à la réalisation sur un axe antérieur/postérieur et le troisième (F3) au caractère étiré/arrondi de la voyelle. La fréquence fondamentale (f0), c'est-à-dire la fréquence de vibration des cordes vocales, détermine la hauteur du son. Elle permet essentiellement de distinguer les grandes classes de sons (voisés/non-voisés), mais donne aussi des informations quant au genre du locuteur, à la structuration temporelle du discours et à l'accentuation. Quant à la durée, il a été montré qu'elle peut être associée au style de parole ou bien à l'appartenance des mots à une classe lexicale ou fonctionnelle (Adda-Decker, 2006).

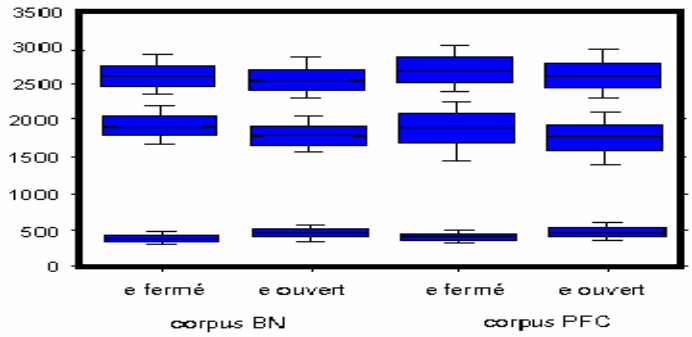

FIG. 1 – *Dispersion des valeurs des trois premiers formants (boîtes à moustache) pour les phonèmes /e/ et /ɛ/ dans le corpus BN et dans le corpus PFC-entretien.*

Pour chaque segment aligné, correspondant à l'un des 4 mots cibles et leur contexte, les mesures sont effectuées toutes les 5*ms*. Un segment comprend ainsi au moins 6 points de mesures, étant donné que la durée minimale d'un segment est de 30*ms*. Pour chaque segment un taux de voisement peut être calculé, correspondant au rapport donné par le nombre de points de mesure avec f0>0 sur le nombre total de points de mesure. La détermination automatique des valeurs de formant est sujette à erreurs. Afin d'éliminer des erreurs de mesure de formants, un filtrage a été mené: un segment est retenu à condition que son taux de voisement n'y soit pas nul.

Pour le mot *est*, il y a la prononciation canonique /ɛ/ (e ouvert) et sa variante /e/ (e fermé) dans le dictionnaire de prononciation utilisé pour la transcription automatique. Ces deux phonèmes sont perceptivement ainsi qu'acoustiquement différents (*cf*. Fig. 1).

2.4 Analyse des paramètres : durée, fréquence fondamentale et cooccurrence de pauses (gauche/droite)

2.4.1 Durée

Le nombre d'occurrences des mots *et/est* et *à/a* a été mesuré en tenant compte du fait que *est* peut être prononcé avec deux timbres de voyelle différents. Dans le corpus BN, il y a 19k occurrences pour le mot *et* et 14k occurrences pour le mot *est* (dont /e/ 9k occ. /ɛ/ 5k occ.). Dans le corpus PFC-entretien le mot *et* est relativement moins fréquent avec 2k occurrences, et 3k occurrences pour le mot *est* (dont le phonème /e/ a 2k occurrences et le phonème /ɛ/ a 1k occurrences). Concernant la paire de mots *à/a, d*ans le corpus BN, on trouve 20k occurrences pour le mot *à* et 11k pour le mot *a*. Quant au corpus PFC-entretien, il y a 4k occurrences pour le mot *à,* contre 3k pour le mot *a*.

Ci-dessous, les courbes détaillent les distributions des mots analysés selon leurs durées respectives. La distribution de la paire de mots *et/est* pour des durées allant de 30*ms* jusqu'à 200*ms* est montrée dans la Figure 2 (corpus BN à gauche et corpus PFC-entretien à droite). Pour le mot *est*, trois courbes sont présentées : la première courbe (losanges) illustre la prononciation majoritaire /e/ (variante de prononciation), la deuxième (carrés) – le phonème /ɛ/ (prononciation canonique), tandis que la troisième courbe (triangles) réunit les distributions des deux prononciations possibles. Chacune des courbes représentées somme à 100%.

Une première tendance qui se dégage est que la durée peut être mise en relation avec le style de parole : les mots sont plus courts dans le corpus PFC (maximum à 30*ms*) que dans le corpus BN (maximum autour de 60-70*ms*). Pour ce qui est des différences entre les homophones *et* et *est*, on observe que la conjonction *et* (courbes rouges à croix) a tendance à avoir une durée plus importante que le verbe. L'intersection des deux mots *et*/*est* est à 80*ms* pour les deux styles de corpus. Après 80*ms*, le taux de *et* est plus important que celui du mot *est*. L'évolution des trois courbes du mot *est* est similaire dans les deux corpus.

La Figure 3 montre des décomptes similaires pour les mots *à* (préposition) et *a* (verbe *avoir*) pour le corpus BN (à gauche) et le corpus PFC-entretien (à droite). Par rapport à la paire de mots *et/est* (*cf*. Fig. 2), la paire *à/a* observe la même tendance de durée plus courte pour le style de parole spontané (PFC-entretien). La préposition *à* est légèrement plus longue que le verbe *a*. Par exemple on peut remarquer dans la Figure 3 à droite (PFC-entretien), qu'un écart d'environ 8% existe pour les deux items *à/a* pour la durée minimale de 30*ms*.

Pour conclure sur cette partie, nous avons analysé les durées de deux paires de mots (quasi-)homophones. Une différence qui semble s'imposer pour les deux paires de mots et les

deux styles de corpus est que la durée de mots-outil de types conjonction ou préposition (surtout *et,* dans une moindre mesure *à*) est plus longue que celle des verbes auxiliaires *est* et *a*. Cette information contribuera éventuellement à différencier nos paires de mots.

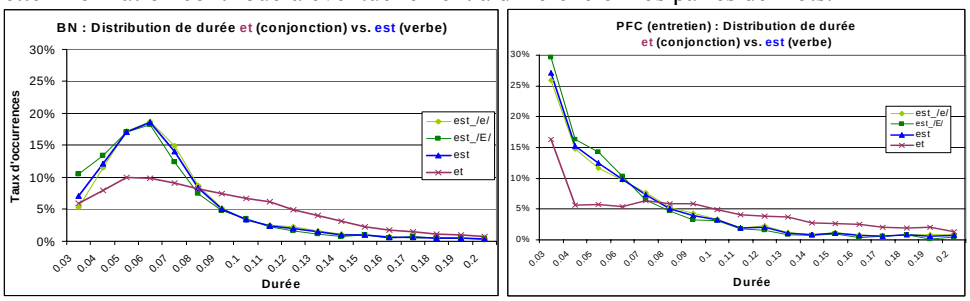

FIG. 2 – *Distributions de durée de « et (conjonction) » et « est (verbe) » du corpus BN (gauche) et PFC (droite).*

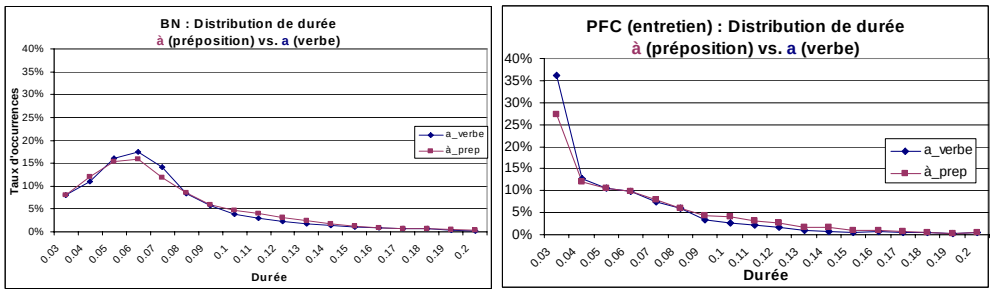

FIG. 3 – *Distributions de durée de « à (préposition) » et « a (verbe) » du corpus BN (gauche) et PFC (droite).*

2.4.2 Fréquence fondamentale (f0)

Les paramètres prosodiques généralement considérés pour caractériser un mot sont la f0, la durée et l'intensité. D'après Selkirk (1996), les mots peuvent être distingués selon la catégorie grammaticale à laquelle ils appartiennent, e.g. fonctionnelle (le déterminant, la préposition, l'auxiliaire, le complément, la conjonction et la particule, etc.) ou lexicale (le nom, le verbe et l'adjectif). Ce statut (fonctionnel vs. lexical) peut être mis en relation avec le paramètre appelé fréquence fondamentale. En effet, on pourrait émettre l'hypothèse qu'un verbe se trouvant à l'intérieur d'un mot prosodique réalise un f0 moyen différent de celui d'une préposition ou d'une conjonction se trouvant en début de mot prosodique. De la même manière, on pourrait penser qu'en début de mot prosodique on peut observer des zones de voisement partielles, précédées éventuellement de césures ou de pauses.
Nous avons donc analysé la f0 en tant qu'information prosodique pour nos paires de mots *et*/*est* et *à*/*a*. La question s'est posée si le taux de voisement (i.e. la proportion de valeurs non nulles de f0) joue un rôle important dans l'articulation des mots homophones analysés. Ainsi, selon le degré de voisement des segments vocaliques correspondant aux mots cibles, les données ont été divisées en trois classes :

1. *Pas voisé* : taux de voisement de 0 à 20% ;
2. *Partiellement voisé* : taux de voisement de 20 à 80% ;
3. *Voisé* : taux de voisement de 80 à 100% .

La première classe (p*as voisé*) ne devrait recueillir qu'une très faible partie des données, sachant que nos analyses se limitent à des mots monophonématiques contenant seulement une voyelle.

Les Figures ci-dessous montrent sous forme d'histogramme le taux d'occurrences des trois classes de voisement, et sous forme de courbe la valeur moyenne de f0. La comparaison des mots *et*/*est* est montrée dans la Figure 4 pour le corpus BN (à gauche) et pour le corpus PFC-entretien (à droite). De gauche à droite, le mot *et* est représenté en prune, le verbe *est* en bleu, /e/ du verbe *est* en vert clair, /ɛ/ du verbe *est* en vert foncé. Le taux d'occurrences de la classe *pas voisé* est très faible pour BN, mais plus important pour PFC-entretien. Cet effet est à mettre en relation avec le style de corpus : nous avons déjà observé les durées très courtes pour le style spontané, et on peut faire l'hypothèse que la parole spontanée soit caractérisée par une hypo-articulation avec élision de voyelles, pouvant entraîner, entre autre, un taux de voisement plus bas.

Dans le groupe appelé *pas voisé*, le taux est un peu plus haut pour le mot *est*. Par contre dans le groupe des occurrences *partiellement voisées*, la tendance est inverse. La conjonction *et* est mieux représentée d'environ 15% que le verbe *est*. Cette situation concerne les deux styles de corpus. Enfin dans le groupe *voisé*, le taux du verbe *est* est plus haut d'environ 10% que celui de la conjonction. On peut ainsi remarquer que globalement le verbe est plus souvent totalement voisé que la conjonction. Si on compare les deux styles de corpus, on remarque que le taux général de voisement est plus bas dans le corpus PFC-entretien.

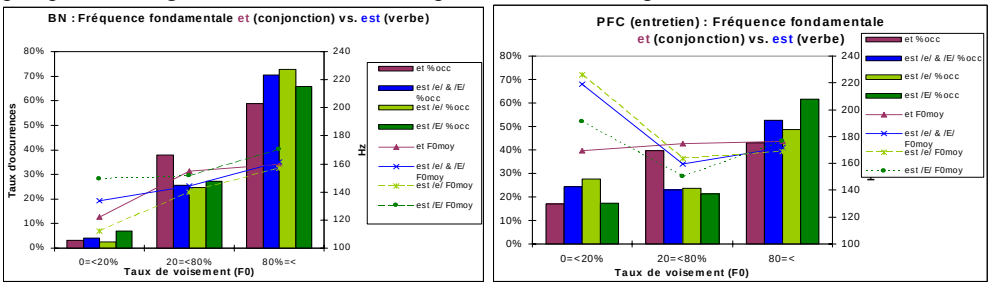

FIG. 4 – *Distribution de pourcentage d'occurrences et valeurs moyennes de* f0 *de « et (conjonction) » et « est (verbe dont la réalisation de prononciation est soit /e/ soit /ɛ/ soit un ensemble de deux) » selon le taux de voisement du corpus BN (gauche) et PFC (droite).*

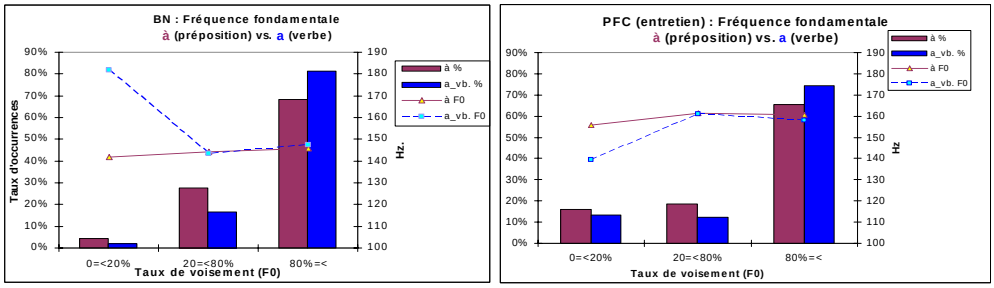

FIG. 5 – *Distribution de pourcentage d'occurrences et valeurs moyennes de* f0 *de « à (préposition) » et « a (verbe) » selon le taux de voisement du corpus BN (gauche) et PFC (droite).*

La comparaison des mots *à*/*a* est montrée dans la Figure 5 pour le corpus BN (à gauche) et pour le corpus PFC-entretien (à droite). De gauche à droite, le mot *à* est illustré en prune,

le verbe *a* en bleu. Comme pour la conjonction *et*, la préposition *à* est plus susceptible d'apparaître dans les groupes à voisement partiel. De manière réciproque, le verbe *a* est mieux représenté dans la classe *Voisé* (les paquets d'au moins 80% de voisement), même si cette tendance est moins forte que pour la paire *et/est*. Cela nous informe donc que dans les deux types de corpus et pour les deux paires de mots, le verbe est plus fréquemment complètement voisé que la préposition ou la conjonction. Le taux de mots dans la classe *Voisé* est plus faible dans le corpus PFC (entretien) que dans le corpus BN, ce qui peut être à nouveau mis en lien avec le style de parole spontanée.

À partir de l'analyse de la f0 et du taux de voisement, on peut remarquer que les verbes auxiliaires sont en général plus voisés que la préposition et la conjonction faisant partie de la catégorie fonctionnelle. Ces résultats indiquent qu'en français, les mots dans la catégorie lexicale pourraient être plus accentués que ceux dans la catégorie fonctionnelle, confirmant ainsi ce qui a été mis en évidence en anglais (Selkirk 1996). Ces mesures permettent d'envisager des attributs permettant de distinguer nos paires de mots homophones. Dans la section suivante nous nous intéressons à ces paires de mots en contexte, afin de trouver des descripteurs distinctifs supplémentaires.

2.4.3 Cooccurrence de pauses (gauche/droite)

Selon Beaugendre et Lacheret-Dujour (1999), les pauses jouent un rôle très important dans le processus d'extraction automatique d'informations prosodiques, et ceci est plus particulièrement vrai pour ce qui est de la parole spontanée. Quant à la perception humaine, les pauses permettent entre autre, de repérer différents contours dans la substance verbale. Nous nous sommes intéressées au rapport qui existe entre les pauses au sens large (qu'il s'agisse d'un silence, d'une pause remplie, i.e. une hésitation, ou d'une respiration) et les paires de mots homophones analysés ici. On examine leurs cooccurrences à **gauche** et à **droite** du mot cible.

Mots	et (conjonction)		est (verbe *être*)		à (préposition)		a (verbe *avoir*)	
Corpus	BN	PFC	BN	PFC	BN	PFC	BN	PFC
Pause gauche	49%	58%	9%	12%	23%	17%	11%	6%
Pause droite	7%	17%	5%	10%	3%	10%	6%	11%

TAB. 1 – *Pourcentage d'occurrences de pauses (silence, respiration, hésitation) à gauche et à droite des mots cibles.*

Le tableau 1 présente le pourcentage d'occurrence de phénomènes listés ci-dessus et regroupés en deux catégories (*pause gauche* et *pause droite*) en comparant les deux paires de mots homophones et les deux styles de corpus BN et PFC entretien.

On peut remarquer que les taux de la catégorie « pause gauche » pour la conjonction *et* sont plus importants que ceux pour le verbe *est*, et ceci dans les deux styles de corpus. Cela suggère que de manière générale, les locuteurs introduisent une césure plus souvent devant la conjonction que devant le verbe et ceci d'autant plus lorsqu'il s'agit de parole conversationnelle (PFC). Pour ce qui est de la paire de mots *à* vs. *a*, des différences comparables sont observées: les pauses sont plus nombreuses devant le mot outil que devant le mot lexical mais dans un degré moindre.

Pour finir, notons donc que la principale différence entre mots fonctionnels (*à*, *et*) et mots lexicaux (*a, est*) concerne l'occurrence de pauses surtout à gauche (et plus légèrement à droite) du mot cible. Ainsi, de manière générale, les verbes *est* et *a* sont rarement précédés d'une pause, au contraire des mots fonctionnels.

2.5 Discussion

À l'issue de cette section, on peut observer que les caractéristiques acoustiques et prosodiques des deux paires de mots homophones établies à partir de dizaines d'heures de parole et quelques milliers d'occurrences pour chaque mot examiné, présentent quelques différences. Des paramètres tels que la durée et le taux de voisement permettent de les distinguer au moins partiellement. La cooccurrence de pauses à gauche et à droite des mots cibles change aussi selon la nature fonctionnelle ou lexicale des homophones considérés. Ainsi, on peut s'interroger si ce type d'attributs acoustiques ne pourrait pas être utilement exploité dans la discrimination de telles paires de mots.

Partant de là, nous nous sommes posé comme objectif de définir des attributs qui pourraient caractériser les mots à travers des techniques de fouille de données. Dans la section suivante, nous décrivons d'abord les descripteurs acoustiques et prosodiques mis en oeuvre, avant d'aborder la méthode adoptée pour discriminer les paires d'homophones analysées ici.

3 Classification des mots homophones par fouille de données

Un ensemble de tests de classification automatique a été mené, visant à déterminer à la fois l'algorithme de classification et les attributs acoustiques les mieux adaptés à distinguer les deux paires de mots homophones *et/est* et *à/a*. Lors de cette étude préliminaire, nous avons fait appel au logiciel Weka (développé à l'université de Waikato, en Nouvelle-Zélande) pour classifier automatiquement ces mots grâce à quelques descripteurs acoustiques et prosodiques pressentis. Le logiciel Weka est destiné à résoudre une variété de problèmes rencontrés dans le cadre de la fouille de données, et le système est équipé d'une large gamme d'algorithmes d'apprentissage et de classification.

3.1 Définition d'attributs

Pour la classification automatique, 41 attributs acoustico-prosodiques ont été définis. Ils ont été choisis pour modéliser à la fois le mot cible (**attributs intra-phonème**) et sa relation au contexte (**attributs inter-phonème**). Ces attributs sont :

Attributs intra-phonème (33) : durée, f0 (moyenne/segment, début, milieu, fin), les formants (F1, F2 et F3) (valeurs moyennes par segment ainsi qu'en début, milieu et fin de segment), taux de voisement (moyennes par segment ainsi qu'en début, milieu et fin de segment). Nous avons également calculé les différences (notées Δ) début-milieu, milieu-fin et début-fin pour les formants et pour la f0.

Attributs inter-phonème (8) : durée, f0, pause. Le paramètre durée est mesuré ici comme suit : la différence entre la durée au centre du segment correspondant au mot cible et le centre de la voyelle précédente/suivante, même s'il y a des consonnes ou des pauses entre ces phonèmes. Pour la f0 au niveau inter-phonémique, Δf0 a été calculée comme différence entre la valeur moyenne de f0 du phonème du mot cible et celle de la voyelle précédente et suivante, et entre ces deux voyelles précédant et suivant le mot cible. Les paramètres « pause gauche » et « pause droite » ont été également rajoutés.

3.2 Expériences de classification

Pour classifier automatiquement les mots à partir de ces attributs, nous avons testé 17 algorithmes implémentés dans le logiciel Weka (classification bayésienne, arbres, règles et fonction etc.), avec le but de trouver l'algorithme le plus performant. Les expériences de classification sont effectuées à l'aide de la méthode de validation croisée, comprenant la division du corpus dans une partie « entraînement » et une partie « test ». Les résultats obtenus sur la partie « test » sont évalués en termes de classification correcte et de coefficient kappa[1]. En fonction des scores de classification et du coefficient kappa, les meilleurs algorithmes sont sélectionnés pour chaque paire de mots. Le tableau 2 ci-dessous montre l'algorithme ayant permis la meilleure discrimination de chaque paire et la moyenne des 5 meilleurs algorithmes/paire de mots. Les résultats montrent que la paire *et*/*est* est nettement mieux classifiée que la paire *à/a*. Cela va dans le sens des résultats de la section précédente où l'on observait que la paire *et/est* se distinguait mieux que la paire *à/a*. Cela s'explique également en partie par le fait qu'un tiers environ des occurrences du verbe *est* ne sont pas de vrais homophones (prononciation /ɛ/ pour *est*) de la conjonction *et* ce qui engendre des attributs plus discriminants. Les résultats pour *et/est* sont particulièrement prometteurs pour le corpus PFC. La parole spontanée présente en général plus d'erreurs lors de la transcription automatique, ces résultats montrent cependant que les homophones analysés sont en grande partie distinguables dans ce type de parole.

Mots	et vs. est				à vs. a			
Corpus	BN		PFC		BN		PFC	
	%	Kappa	%	Kappa	%	Kappa	%	Kappa
LMT	78	0.55	95	0.90	72	0.35	66	0.32
Meilleur parmi les 17 algorithmes	78	0.55	97	0.93	72	0.35	66	0.32
Moyenne sur 5 meilleurs algo.	77	0.53	96	0.91	72	0.35	65	0.30

TAB. 2 – *Taux de classification correcte des mots par les algorithmes testés dans Weka.*

3.3 Sélection d'attributs

41 attributs ont été utilisés pour classifier les paires de mots. Nous pourrions faire l'hypothèse que parmi ces attributs, certains sont plus pertinents que d'autres. Les 10 meilleurs attributs ont été ainsi retenus (*cf.* tableau 4) à partir des résultats données par l'algorithme le plus performant, i.e. l'algorithme LMT (*Logistic Model Trees*). À partir de ces 10 attributs sélectionnés, nous avons recalculé les pourcentages de classification correcte avec l'algorithme LMT en utilisant la méthode de la validation croisée. Les attributs sélectionnés et leurs résultats sont présentés dans le tableau 3. Les résultats à partir de 10 attributs sont légèrement moins performants que l'utilisation de tous les attributs, surtout pour la paire *et/est* dans le corpus PFC. Le résultat du corpus BN montre que la différence entre 10 attributs et 41 attributs reste marginale (1%). Ceci montre que seuls 10 attributs suffisent pour

[1] Le coefficient kappa mesure ici la concordance entre la classification automatique par Weka et les deux classes réelles de paramètres caractérisant les mots *et/est*, *à/a*. Il varie entre -1 (désaccord total) et 1 (accord total), en passant par 0 (classification au hasard). En fonction des résultats de la classification, les seuils correspondant aux meilleurs algorithmes sont : k>0.50 pour *et/est* et k>0.25 pour *à/a*. Cela montre d'emblée une différence entre les deux paires : le kappa de *et/est* est très bon tandis que les résultats moins bons pour *à/a* montrent que cette paire est beaucoup plus difficile à discriminer.

produire des résultats pratiquement équivalents à ceux obtenus avec 41 attributs. Il y a donc des traits acoustiques et prosodiques particulièrement significatifs qui permettent de distinguer les mots homophones et ces traits sont encodés en grande partie dans les 10 attributs.

Mots	et vs. est		à vs. a	
Corpus	BN	PFC	BN	PFC
LMT 10 attributs	77% (0.53)	89% (0.78)	71% (0.32)	67% (0.34)
LMT 41 attributs	78% (0.55)	95% (0.90)	72% (0.35)	66% (0.32)

TAB. 3 – *Comparaison des taux de classification des mots (en %, coefficient kappa entre parenthèses) en fonction du nombre d'attributs (10 vs. 41) avec l'algorithme LMT.*

Mots	*et* vs. *est*		*à* vs. *a*	
Corpus	BN	PFC	BN	PFC
1	**pause G**	*ΔF2 déb.-fin*	**Δdurée D**	**Δdurée D**
2	**Δdurée G**	*durée*	*F2 milieu*	**Δdurée G**
3	*ΔF1 déb-milieu*	**Δdurée G**	*F2 début*	**pause G**
4	**Δdurée D**	*F2 milieu*	**pause G**	*Δf0 déb.-milieu*
5	**Δf0 D**	**pause G**	**Δf0 G**	*ΔF2 déb.-milieu*
6	*durée*	**Δdurée G-D**	*f0 déb. voisement*	*f0 déb. voisement*
7	**Δf0 G**	**pause D**	*durée*	**Δf0 G**
8	*f0 voisement*	*F2*	*F2 fin*	*F2 début*
9	**Δf0 G-D**	*ΔF1 déb.-milieu*	**Δdurée G**	**Δf0 G-D**
10	**Δdurée G-D**	*F3 début*	**pause D**	*F1 début*

TAB. 4 *–10 attributs (intra-segmentaux en italiquer et inter-segmentaux en gras) mieux classés par l'algorithme LMT.*

En regardant de plus près les attributs les plus discriminants, un certain nombre de tendances se dégage. En ce qui concerne la paire *et*/*est*, on remarque que les paramètres concernant les durées (durée du phonème, Δdurée inter-phonèmes) et l'existence d'une pause devant le mot cible s'avèrent particulièrement discriminants. Les paramètres liés au voisement (taux de voisement du phonème, Δf0 inter-phonèmes) pour le corpus BN et aux formants (F2, F2 milieu, F3 début, ΔF1, ΔF2) pour le corpus PFC-entretien sont aussi discriminants. Les paramètres inter-phonémiques sont surtout pertinents pour le corpus BN.

Pour la paire *à*/*a*, on peut observer d'abord que les paramètres inter-phonémiques sont importants : ceux liés aux durées (Δdurées inter-phonèmes) et au voisement (Δf0 inter-phonémes). Ensuite, le paramètre « pause gauche » est également utile. Enfin, les paramètres concernant le deuxième formant (F2 début, F2 milieu, F2 fin, ΔF2), ainsi que le voisement jouent un rôle non négligeable. Il est aussi à noter qu'il y a beaucoup de paramètres communs aux deux corpus qui permettent de distinguer les deux mots.

Cette analyse, bien que préliminaire, a permis de mettre en évidence que les mots homophones peuvent être différenciés grâce à certains paramètres acoustiques et prosodiques. Ce fait est valable pour les deux paires et surtout pour *et/est* dans les deux corpus. Les résultats obtenus suggèrent des pistes intéressantes pour mieux traiter ces mots lors de la transcription automatique de la parole. Afin de valider cette approche et les paramètres retenus, l'analyse devrait s'étendre à d'autres mots homophones et d'autres types de corpus.

4 Conclusion et perspectives

Dans ce travail nous avons cherché à étudier les réalisations acoustiques des mots homophones dans des grands corpus oraux de différents styles de parole (parole préparée, parole

conversationnelle) avec le but à plus long terme de pouvoir contribuer à la modélisation acoustique des variantes de prononciation dans les systèmes de transcription automatique de la parole. Notre objectif était ensuite d'examiner les réalisations acoustiques de mots homophones fréquents, sources de nombreuses erreurs de transcription, de définir et d'évaluer des attributs acoustiques, permettant éventuellement de classifier correctement ces homophones. À cet effet nous avons utilisé différents outils (STK-LIMSI pour l'alignement automatique, Praat pour l'extraction d'attributs acoustiques, Weka pour la classification et la fouille de données).

Nous nous sommes servies de l'expérience acquise sur ces grands corpus de parole variés, pour définir des attributs acoustiques et prosodiques potentiellement utiles pour la discrimination des mots, en particulier des mots (quasi-)homophones. Ainsi nous avons pu étudier plus particulièrement les paires de mots outils homophones, *et*/*est* et *à*/*a*, qui sont les couples de mots qui produisent le plus d'erreurs lors de la transcription automatique. La comparaison de durées entre *et* (conjonction) /*est* (verbe *être*) montre que le mot *est* a tendance à être réalisé avec une durée beaucoup plus faible, alors que la conjonction *et* se trouve souvent allongée. Cette simple mesure suggère que les homophones, réalisés a priori avec les mêmes phonèmes (par exemple, les mêmes valeurs de formants pour les voyelles), peuvent différer dans leur réalisation prosodique. Par la « réalisation prosodique » nous entendons tout ce qui concerne durée, fréquence fondamentale, voisement des segments, ainsi que leur articulation avec les contextes droite et gauche, incluant des mesures de pauses. On aboutit ainsi à un ensemble de mesures intra- et inter-phonèmes, servant à définir les 41 attributs acoustiques utilisés pour la classification. Pour les expériences de classification, nous avons testé les différents algorithmes proposés dans Weka : classification par arbres de décision, classification par règles, classification par MVS (machines à vecteurs support), classification bayésienne. Le résultat de la classification automatique par les 17 algorithmes testés dans le logiciel de fouille de données montre que les attributs impliquant les durées intra- et inter-segmentales de la paire *et*/*est* sont parmi les plus importants pour la classification.

La classification automatique s'est avérée particulièrement prometteuse pour la paire *et/est* : les résultats sont supérieurs à 70% (BN) voire 90% pour le corpus PFC. Les mots *à/a* sont discriminables aussi, même si de manière moins aisée que le paire *et/est*. Ces résultats montrent qu'il existe des informations acoustiques pertinentes pour la discrimination des homophones. Leur utilisation explicite dans les systèmes de transcription devrait contribuer dans le futur à réduire le taux de confusions observées.

Nous rappelons ici les attributs qui se sont avérés les plus utiles pour la classification: les durées inter-segmentales (en lien avec un phonème vocalique précédent ou/et suivant) sont plus efficaces que la durée du segment propre. La même remarque vaut également pour la f0.

Dans des travaux futurs, nous allons essayer de mieux caractériser les homophones en rajoutant de nouveaux attributs pour la discrimination, et ensuite tenter d'implémenter efficacement les attributs identifiés (en particulier mesures de durée et de f0 inter-segmentales) pour améliorer la transcription automatique. Nous comptons également, dans le futur, étendre ce type d'études à plus de mots, et intégrer alors des informations morpho-syntaxiques, afin de mieux factoriser les variantes observées dans la parole.

Remerciements

Les auteurs tiennent à remercier Bianca Vieru-Dimulescu pour son aide avec l'outil Weka. Les travaux présentés ont été partiellement financés dans le cadre des projets *AMADEO* du RTRA DIGITEO et .ANR PFC-Cor

Références

Adda-Decker, M. (2006). De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux. In Journées d' Étude sur la Parole, Dinard, Juin.

Beaugendre, F. et Lacheret-Dujour, A. (1999). *La prosodie du français*. CNRS Langage, avril.

Boersma, P. and Weenink, D. (1999-2007). Praat: doing phonetics by computer, www.praat.org.

Durand, J., Laks, B. et Lyche, C. (2003). Le projet "Phonologie du français contemporain (PFC)", La Tribune International des Langues Vivantes 33 : 3-9.

Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.-F. and Gravier, G. (2005). The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In *EUROSPEECH*, Lisbonne.

Gauvain, J.L., Adda, G., Adda-Decker M., Allauzen A., Gendner V., Lamel, L. and Schwenk, H. (2005) Where Are We in Transcribing French Broadcast News? In *Inter-Speech*, Lisbon, September.

Selkirk, E. (1996). The prosodic structure of function words. In *Signal to syntax: bootstrapping from speech to grammar in early acquisition*, e.d. J. L. Morgan & E. Demuth, pages 187-214, Lawrence Erlbaum Associates, Mahwah.

Vieru-Dimulescu B., Boula de Mareüil P., Adda-Decker M. (2007). Identification of foreign-accented French using data mining techniques. In *ParaLing07*, Saarbrücken

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco.

Summary

This article represents a contribution to the acoustic modeling of words' pronunciation with the aim of factorizing the observed word variants in speech. A significant number of automatic speech transcription errors are concerned with frequent homophone words. Here two (quasi-)homophone word pairs in French, i.e. *à*/*a* (preposition (to)/verb (*have*, *take*)) and *et*/*est* (conjunction (*and*)/ verb (*be*)) are chosen to examine acoustic and prosodic attributes which potentially allow distinguishing homophone word pairs using data mining techniques. Databases illustrating two different speaking styles (e.g. prepared *ESTER-BN* vs. spontaneous speech *PFC interview*) are employed to this purpose. 41 acoustic and prosodic attributes are then defined and used to automatically discriminate the two pairs. Automatic classification gives an average of 77% using the 5 best algorithms. When reducing the number of descriptors to 10 (selected thanks to the most efficient algorithm), the automatic classification results are close to the ones obtained with 41 attributes suggesting some acoustic/prosodic attributes are particularly salient for homophone words discrimination.