Mots fréquents homophones en français : analyse acoustique et classification automatique par fouille de données

Rena Nemoto, Ioana Vasilescu, Martine Adda-Decker

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

ABSTRACT

Many automatic speech transcription errors arise from frequent homophone words. This paper aims at verifying whether acoustico-prosodic attributes contribute to discriminate homophones without higher level linguistic information. We chose two homophone pairs, i.e. *à*/*a* (*to*/*has*) and *et*/*est* (*and*/*is*) from two different speech (prepared vs. spontaneous) corpora. Acoustic analyses show differences in voice ratios and duration between verbs and conjunction/preposition homophones. Combining pitch, formants, duration, voice ratio and cooccurring pause measures, 62 acoustic-prosodic features were defined for classification. Average identification rates between 60 and 77% have been achieved. Major features are of prosodic and contextual inter-phonemic nature.
Keywords : homophone, French, acoustic-prosodic features, voice ratio, data mining, classification.

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 d'effectuer des analyses acoustiques à grande échelle afin d'extraire des connaissances relatives aux spécificités acoustiques et prosodiques caractérisant les prononciations de mots homophones. Ces connaissances peuvent servir à définir des attributs pour la classification automatique par fouille de données. Cette approche a déjà pu montrer son intérêt pour la caractérisation des accents étrangers [8] et des premiers résultats concernant les homophones sont décrits dans [7]. Nous nous intéresserons ici aux mots considérés comme homophones, par exemple un verbe au participe passé ou à l'infinitif (*allé*, *aller*), qui sont facilement confondu lors de la transcription automatique. Partant de ces constats, nous nous sommes interrogées si les mots homophones ne déploieraient pas de particularités acoustiques/prosodiques qui n'ont pas été prises en compte ni par les paramètres acoustiques classiques (cepstres), ni par les modèles acoustiques (Modèles de Markov Cachés) et qui permettraient leur discrimination. Nous faisons ainsi l'hypothèse que des informations prosodiques (durée, fréquence fondamentale notée f0, intensité, 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). Dans cette article, deux paires de mots *et* (conjonction)/*est* (verbe *être*) et *à* (préposition)/*a* (verbe *avoir*), sont choisies pour définir et examiner des descripteurs ou attributs acoustiques et prosodiques. Ces mots sont parmi les plus fréquents du français et souvent mal reconnus lors de la transcription automatique [1]. 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* correspond à /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. Nous avons fait appel aux techniques de fouille de données afin de classifier automatiquement ces mots grâce à un ensemble d'attributs acoustico-prosodiques pour ce travail.
Dans la section 2, nous allons présenter les corpus utilisés et la section 3 décrit les analyses menées concernant la durée et le taux de voisement. 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 4), avant de conclure et d'ouvrir quelques perspectives (section 5).

2. Corpus

Cette étude part de deux types de corpus en français : 55 heures de journaux radiodiffusés de différentes sources (France Inter, RFI, France Info et RTM) collectés pour la campagne ESTER (Evaluation des Systèmes de Transcription enrichie d'Emissions Radiophoniques) [4], appelé **BN** (*Broadcast News*). Le style y est (semi-)préparé. L'autre corpus, appelé **PFC** (Phonologie du Français Contemporain) [3] contient des enregistrements de variétés de français de régions différentes et en différents styles de parole. Nous avons retenu un sous-ensemble de 20 heures composé d'entretiens correspondant à l'oral spontané (avec 12 heures de parole effective).

2.1. Alignement automatique

Le système de transcription automatique de parole du LIMSI [5] est utilisé pour l'alignement des corpus décrits ci-dessus. L'alignement utilise des transcriptions manuelles, un dictionnaire de prononciation et des modèles acoustiques de phones indépendants du contexte. Il localise les mots prononcés et les pauses,

détermine leur prononciation et segmente le flux audio en phones, permettant ainsi de mesurer la durée de nos deux paires de mots, ainsi que de faire des mesures contextuelles. Le tableau 1 montre le nombre d'occurrences des mots *et*/*est* et *à*/*a* en tenant compte du fait que *est* peut être prononcé avec deux timbres de voyelle différents (les prononciations incluant la liaison ([ɛt], [et]) ont été écartées).

Tab. 1: Nombres d'occurrences de mots.

	BN		PFC	
à	20,4k	/a/	3,6k	/a/
a	11,3k	/a/	3,4k	/a/
et	19,1k	/e/	5,0k	/e/
est	14,5k	/ɛ/5,0k, /e/9,5k	6,2k	/ɛ/1,9k, /e/4,3k

3. Analyses acoustiques

Le logiciel PRAAT [2] 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. Nous avons ainsi extrait les trois premiers formants (pour plus de détails, voir [7]), la fréquence fondamentale (f0) et l'intensité. Pour chaque segment aligné, les mesures sont effectuées toutes les $5ms$. Un taux de voisement peut être calculé ainsi : une trame est considérée comme voisée dès lors que la f0 y est définie ($Pv = \frac{nombre\ de\ trames\ voisées}{nombre\ de\ trames}$). Afin de minimiser les erreurs de mesure de formants, un filtrage simple a été effectué : un segment n'est retenu que si son taux de voisement n'y est pas nul. Dans les Figures 1 et 2 (partie supérieure), le mot *et* est représenté en rouge et le verbe *est* en bleu (la variante de prononciation [e] en vert clair, la prononciation canonique /ɛ/ en vert foncé). Dans les mêmes figures (en bas), le mot *à* est en rouge et le mot *a* en bleu.

Durée Dans la Figure 1 les courbes détaillent les distributions des mots analysés selon leurs durées respectives allant de $30ms$ jusqu'à $200ms$ (BN à gauche, PFC à droite). Pour faciliter les comparaisons, chacune des courbes représentées somme à 100% de taux d'occurrences. 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 (max. à $30ms$) que dans le corpus BN (max. 60-70ms). Pour ce qui est des différences entre les homophones *et* et *est*, on observe que la conjonction *et* a tendance à avoir une durée plus importante que le verbe. Après $80ms$, le taux de *et* est plus important que celui du mot *est*. L'évolution des trois courbes du mot *est* est similaire dans chacun des deux corpus. La paire *à*/*a* observe la même tendance que la paire *et*/*est*, même si la durée de *à* reste globalement très proche de celle de *a*. La durée de mots-outil (surtout *et*, dans une moindre mesure *à*) est plus longue que celle des verbes. Cette information contribuera éventuellement à différencier nos paires de mots.

Fréquence fondamentale (f0) Les paramètres prosodiques généralement considérés sont la f0, la durée et l'intensité. Les mots considérés ici correspondent à des parties du discours différentes : verbe (*est*, *a*), conjonction de coordination (*et*) et préposition (*à*). Ces mots occupent a priori des positions différentes dans la phrase. 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, et assumant de ce fait un rôle démarcatif. Ainsi, on pourrait penser qu'en début de mot prosodique le voisement du segment ne soit pas total, et qu'il se trouve précédé éventuellement de césures ou de pauses. La question s'est posée si le taux de voisement joue un rôle important dans l'articulation des mots analysés. Ainsi, selon le degré de voisement des segments cibles, les données ont été divisées en trois classes :
1. *Pas voisé* : % de voisement de 0 à 20% ;
2. *Partiellement voisé* : % de voisement de 20 à 80% ;
3. *Voisé* : % de voisement de 80 à 100%.

Dans la Figure 2, le taux d'occurrences des trois classes de voisement est montré en histogramme. La comparaison des mots *et*/*est* est montrée en haut et celle des mots *à*/*a* en bas. On peut remarquer quelques tendances communes pour les deux paires. Comme attendu, la catégorie *pas voisée* renferme une très faible partie de données. Dans le groupe *partiellement voisé*, les taux de voisement de la conjonction et de la préposition (*et* et *à*) sont plus élevés que ceux des verbes (*est* et *a*). De manière réciproque, les verbes *est* et *a* sont mieux représentés dans la classe *voisé*. On peut remarquer que les verbes sont en général plus voisés que la préposition et la conjonction. Si on compare ces deux différents styles de parole, le taux d'occurrences de la classe *pas voisé* est très faible pour BN, mais plus important pour PFC. Cet effet est à mettre en relation avec le style de corpus : en effet des durées très courtes sont observées fréquemment pour le style spontané. Ainsi, 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. Ces mesures permettent d'envisager des attributs pour distinguer nos paires de mots homophones.

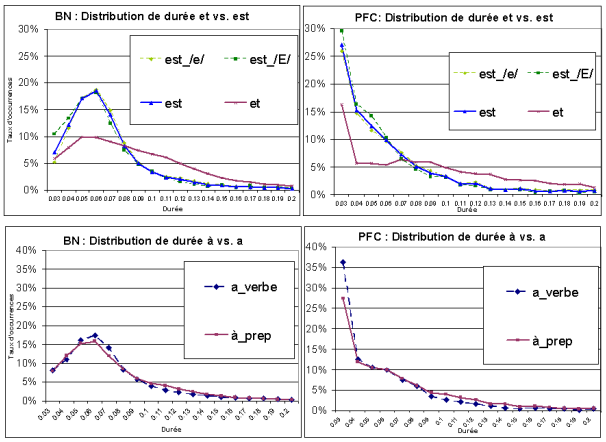

Fig. 1: Distributions de durée des deux paires homophones sur les corpus BN (gauche) et PFC (droite). **En haut :** *et* (rouge), *est* (bleu) (/e/ vert clair, /ɛ/ vert foncé). **En bas :** *à* (rouge), *a* (bleu) .

Cooccurrence de pauses (gauche/droite) Selon Beaugendre et Lacheret-Dujour [6], les pauses jouent un rôle très important dans le processus d'extraction automatique d'informations prosodiques, et ceci est particulièrement vrai pour la parole spontanée. Nous nous sommes intéressées au rapport qui existe entre les pauses au sens large (silence, respiration et pause remplie, i.e. une hésitation) et les homophones ana-

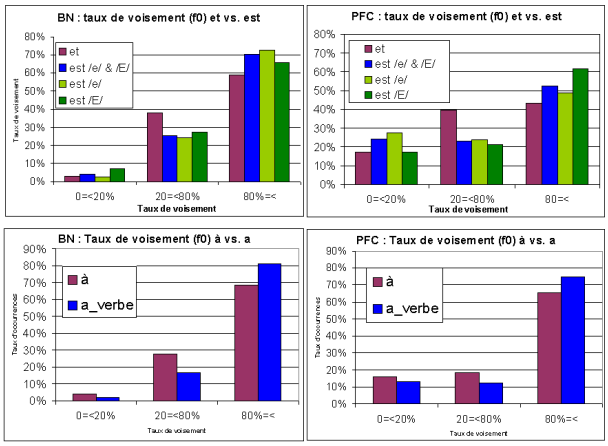

Fig. 2: Distributions des occurrences (en %) des deux paires homophones sur les corpus BN (gauche) et PFC (droite) selon le taux de voisement. **En haut :** *et* (rouge), *est* (bleu) (/e/ vert clair et /ɛ/ vert foncé). **En bas :** *à* (rouge), *a* (bleu).

lysés. On examine leurs cooccurrences à **gauche** et à **droite** du mot cible. Le tableau 2 présente le taux d'occurrence de pauses regroupés en deux catégories (*pause à gauche* et *pause à droite*) en comparant les deux paires de mots homophones et les deux styles de corpus BN et PFC entretien. La principale différence entre préposition/conjonction (*à*, *et*) vs verbes (*a*, *est*) concerne l'occurrence de pauses surtout à gauche du mot cible. Ainsi, de manière générale, les verbes *est* et *a* sont rarement précédés d'une pause.

Tab. 2: % d'occurrences de pauses (silence, respiration, hésitation) à gauche et à droite des mots cibles.

Mots	et		est		à		a	
Corp.	BN	PFC	BN	PFC	BN	PFC	BN	PFC
P.G.	**49**	**58**	9	12	23	17	11	6
P.D.	7	17	5	10	3	10	6	11

4. Classification des mots homophones par fouille de données

On a pu observer que des paramètres tels que la durée, le taux de voisement et la cooccurrence de pauses autour des mots cibles permettent de les distinguer au moins partiellement. Partant de là, nous avons utilisé ces attributs pour vérifier s'ils peuvent aider à discriminer les homophones à travers des techniques de 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 acoustico-prosodiques les mieux adaptés à distinguer les deux paires de mots homophones. Lors de cette étude, nous avons fait appel au logiciel Weka [9].

4.1. Définition d'attributs

62 attributs acoustico-prosodiques ont été définis pour la classification automatique. 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 (40) : durée, f0, taux de voisement, les formants, intensité (valeurs 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 la f0, les formants et l'intensité.
Attributs inter-phonème (22) : durée, f0, les formants, intensité, 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, les formants, et l'intensité au niveau inter-phonémique, Δ a été calculée comme différence entre la valeur moyenne 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 à gauche et pause à droite ont été également rajoutés.

4.2. Expériences de classification

Pour classifier automatiquement les mots à partir de ces attributs, nous avons testé 25 algorithmes implémentés dans le logiciel Weka (classification bayésienne, arbres, règles et fonction etc.). Les expériences de classification sont effectuées à l'aide de la méthode de validation croisée. Le tableau 3 montre l'algorithme ayant permis la meilleure discrimination de chaque paire, la moyenne des 10 meilleurs algorithmes/paire de mots et la moyenne des 25 algorithmes par 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 intéressants pour le corpus PFC puisque la parole spontanée présente en général plus d'erreurs lors de la transcription automatique.

4.3. Comparaison de différents types d'attributs

Nous pourrions faire l'hypothèse que parmi ces 62 attributs, certains sont plus pertinents que d'autres. Nous avons catégorisé ces attributs en quatre classes : prosodie (32 att.), formants (40 att.), inter-segments (22 att.), intra-segments (40 att.).
Les résultats [1] (cf. tableau 3) avec les attributs prosodiques et inter-segmentaux restent presque identiques en comparaison avec les résultats obtenus par 62 attributs pour chaque paire et chaque corpus. Les résultats avec les attributs formantiques et intra-segmentaux sont moins performants, particulièrement pour la paire *et*/*est*. Ce tableau 3 montre que seuls la moitié ou un tiers d'attributs suffisent pour produire des résultats pratiquement équivalents à ceux obtenus avec 62 attributs. Ainsi les traits prosodiques sont particulièrement intéressants pour distinguer les mots homophones, en revanche le timbre de la voyelle

[1] Le coefficient kappa mesure ici la concordance entre la classification automatique et les deux classes réelles de paramètres caractérisant les mots *et*/*est*, *à*/*a*. Il varie entre -1 (désaccord total) et 1 (accord total). Pour *et*/*est*, les valeurs moyennes sur 10 meilleurs algorithmes avec 62 paramètres sont ≥0,50 pour chaque corpus, ce qui va dans le sens des résultats du tableau 3. Pour *à*/*a*, les valeurs sont ≥0,30. Les valeurs de kappa pour *et*/*est* sont très encourageantes tandis que la classification de *à*/*a* s'avère plus difficile.

Tab. 3: Comparaison des % de classification des mots en fonction des types d'attributs par les algorithmes testés dans Weka. Dans le tableau le meilleur%, la moyenne sur 10 meilleurs algorithmes et la moyenne sur 25 algorithmes sont montrés. Le nombre d'attributs pour chaque catégorie est marqué entre parenthèses.

Mots	et vs. est						à vs. a					
Corpus	BN			PFC			BN			PFC		
	meill.	10meill.	moy.	meill.	10meill.	moy.	meill.	10meill.	moy.	meill.	10meill.	moy.
tous (62)	79.8	77.8	71.3	83.1	81.1	76.3	72.9	71.4	66.3	69.4	66.4	61.6
formants (30)	67.5	65.9	**62.3**	66.6	65.3	**62.7**	69.0	67.7	64.3	62.7	61.2	58.5
prosodie (32)	79.5	77.7	**70.9**	82.4	81.0	**77.3**	72.3	70.6	65.6	67.7	65.9	60.7
intra- (40)	73.2	71.3	65.7	71.7	70.4	67.0	68.9	68.0	64.0	60.0	59.3	57.0
inter- (22)	75.7	74.4	69.2	81.2	80.5	77.0	71.0	70.1	65.5	65.9	65.1	60.1

/E/ ne semble pas jouer un rôle prépondérant.

5. Conclusion et perspectives

Dans ce travail nous avons vérifié si les paramètres acoustico-prosodiques, ne contenant pas d'informations contextuelles, pouvaient aider à discriminer les mots fréquents homophones (*et*/*est* et *à*/*a*) dans des grands corpus oraux de différents styles de parole (préparée vs. spontanée) en utilisant les techniques de fouille de données. A cet effet nous avons utilisé différents outils (le système LIMSI pour l'alignement automatique, Praat pour l'extraction d'attributs acoustiques, Weka pour la classification automatique).
Les analyses ont montré que les mots fonctionnels (*et*,*à*) ont un taux de voisement plus faible que les verbes (*est*,*a*) et qu'ils sont plus souvent précédés de pauses. La comparaison de durées entre *et* (conjonction)/*est* (verbe *être*) montre que le mot *est* est souvent réalisé avec une durée faible, alors que la conjonction *et* se trouve souvent allongée. Ces mesures suggèrent 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, f0, voisement des segments, intensité, 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 62 attributs acoustiques utilisés pour la classification. Les résultats de la classification automatique utilisant soit l'ensemble des attributs, soit des sous-ensemble limités aux formants ou à la prosodie, ou centrés sur le segment ou son environnement, montrent que les attributs prosodiques et intersegmentaux sont les plus importants pour la classification. Ceci confirme 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 les taux de confusions observées.

Dans des travaux futurs, nous envisageons d'étendre ce type d'études à plus de mots, et d'intégrer alors des informations morpho-syntaxiques, afin de mieux factoriser les variantes observées dans la parole. D'autres méthodes d'analyse de données pourraient également être considérées (par. ex. analyse factorielle) afin de valider la classification des mots en fonction de leurs particularités acoustiques et prosodiques.

6. Remerciements

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

Références

[1] M. Adda-Decker. De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux. In *JEP*, Dinard, France, 2006.

[2] P. Boersma and D. Weenink. *Praat : doing phonetics by computer*. Institute of Phonetic Sciences, University of Amsterdam, Pays-Bas, 1999–2007. http ://www.fon.hum.uva.nl/praat/.

[3] J. Durand et al. Le projet "Phonologie du français contemporain (PFC)". *La Tribune International des Langues Vivantes*, 33 :3–9, 2003.

[4] S. Galliano et al. The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In *Eurospeech*, Lisbonne, 2005.

[5] J.-L. Gauvain et al. Where Are We in Transcribing French Broadcast News ? In *InterSpeech*, Lisbonne, 2005.

[6] Anne Lacheret-Dujour and Frédéric Beaugendre. *La prosodie du français*. CNRS, Paris, 1999.

[7] R. Nemoto, M. Adda-Decker, and I. Vasilescu. Fouille de données audio pour la classification automatique de mots homophones. In *EGC'2008*, Sophia-Antipolis, France, 2008.

[8] B. Vieru-Dimulescu et al. Identification of foreign-accented French using data mining techniques. In *ParaLing07*, Saarbrücken, 2007.

[9] I. H. Witten and E. Frank. *Data Mining : Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2005.