Diphtongaison et identification automatique dans les dialectes de l'anglais britannique

Emmanuel Ferragne

Laboratoire Dynamique Du Langage – UMR 5596 CNRS – Université Lyon 2

Abstract

Cette étude s'inscrit dans le cadre de nos travaux sur l'identification des dialectes de l'anglais des Iles Britanniques à partir de traits phonético-phonologiques. Il s'agit d'estimer dans quelle mesure la diphtongue de l'ensemble lexical FACE pourrait contribuer à un système de classification automatique des dialectes par la machine. A partir de mesures de pentes des deux premiers formants, nous avons procédé à une classification par le biais d'un réseau de neurones artificiels (perceptron), considérant l'une après l'autre les 91 paires de dialectes de notre corpus. Lorsque les deux dialectes ont un type de voyelle différent du point de vue de la diphtongaison (monophtongue vs. diphtongue, diphtongue fermante vs. diphtongue centripète, etc.), le taux d'identification correcte avoisine les 100%.

Introduction

L'étude des dialectes – ou plus précisément des accents, dans notre cas – peut être conduite de plusieurs manières selon le but recherché. Par exemple, dans le paradigme sociolinguistique/variationniste (e.g Foulkes & Docherty, 1999 ; Kortmann & Schneider, 2004), l'objectif consiste à cerner les facteurs sociaux qui engendrent des variantes de prononciation. Il s'agit très souvent de dialectologie urbaine. Une dialectologie plus

traditionnelle du type de celle qui a servi de cadre au Survey of English Dialects met généralement l'accent sur la conservation de variantes de prononciations archaïsantes et moribondes. Notre approche ne s'oppose pas aux deux que nous venons de mentionner : elle se différencie bien plus sur le plan de la forme que du fond ; il s'agit davantage d'un nouvel outil que d'une véritable rupture épistémologique. En effet, notre étude se caractérise par l'utilisation de technologies actuelles, le but n'étant en réalité déterminé que par le corpus spécifique qui est employé.

Notre objectif est de proposer un système informatique capable de déterminer l'origine géographique d'un locuteur anglophone des Iles Britanniques[1]. A quoi cela peut-il bien servir ? En plus de retombées descriptives indéniables, il a été démontré que les systèmes de reconnaissance de la parole ont de meilleures performances lorsqu'ils sont adaptés au dialecte. Certaines applications judiciaires sont également envisageables. La procédure automatique autorise un gain de temps considérable, ce qui laisse envisager des applications fonctionnant en temps réel, et garantie une reproductibilité totale. Cette étude sur la diphtongaison est motivée par la recherche d'une certaine parcimonie quant aux nombres de paramètres du modèle ; autrement dit, elle privilégie non pas l'approche brute (qui serait celle de l'ingénieur), mais plutôt une approche raisonnée à mi-chemin entre la phonétique traditionnelle et l'ingénierie de la parole.

Nous nous concentrons ici sur la voyelle de l'ensemble lexical FACE. Après avoir décrit le corpus utilisé dans cette étude et les types de voyelles rencontrés, nous exposerons les résultats de notre expérience de classification automatique.

Corpus

Le corpus Accents of the British Isles (ABI) est une base payante d'enregistrements effectués en 2003. Il contient 14 aires dialectales, 284 locuteurs et locutrices en tout, soit en moyenne 20 locuteurs par dialecte (10 hommes et 10 femmes). La partie qui nous intéresse est un passage lu d'environ 200 mots qui a été conçu précisément pour mettre en évidence la variation dialectale. Le corpus est livré avec une transcription orthographique.

Le corpus ABI possède une qualité évidente : sa taille. En revanche, ses défauts, parfois véritablement affligeants, sont nombreux. En premier

[1] Il aurait été tout aussi envisageable de tenter de déterminer l'origine sociale d'un locuteur si des données sociolinguistiques avaient été incluses dans notre corpus.

lieu, l'absence de données sur les locuteurs le rendent inexploitable pour toute étude sociolinguistique. En effet, hormis l'information disponible sur le sexe du locuteur et le fait que l'âge des locuteurs varie de 18 à 60 ans, il est tout à fait impossible de connaître l'âge d'un locuteur précis, sans parler de son activité professionnelle, le quartier dans lequel il a grandi, son niveau d'études, etc. A ce stade, notons qu'il est absolument incroyable que les ingénieurs de l'Université de Birmingham qui ont collecté le corpus ABI n'aient pas eu une conscience de la linguistique plus développée ! Puis, il convient de noter que le texte lu est construit de manière particulièrement maladroite et tend à pousser certains locuteurs à se risquer à un exercice de déclamation épique qu'il aurait été préférable d'éviter lorsque l'on cherche à étudier les sons de la parole tels qu'ils sont réalisés dans des actes de communication naturelle. Notons encore un biais méthodologique imputable à la politique de recrutement des sujets : ceux-ci répondaient à une annonce qui leur offrait d'empocher 10 livres sterling s'ils avaient un accent local… Enfin, certains sujets sont de très mauvais lecteurs (un casse-tête pour le traitement automatique), et les enregistrements ont eu lieu dans des pièces aux propriétés acoustiques très différentes avec, parfois, des bruits externes capturés par le microphone.

Segmentation

La segmentation du corpus en phones a été réalisée à l'aide du Hidden Markov Model Toolkit (HTK)[2]. Il s'agit d'un alignement forcé. Le principe, dans les grandes lignes, est le suivant : le processus comporte une phase d'apprentissage pendant laquelle des valeurs acoustiques typiques pour chaque phone, ainsi que des probabilités d'enchaînements de phones, sont obtenues à partir d'un autre corpus. Ensuite, étant donnée une transcription orthographique correspondant à une portion de signal acoustique donné, le système, fort de ce qu'il a appris lors de la phase précédente, s'applique à repérer au mieux les frontières entre phones.

Description de la voyelle de FACE

D'après les références de dialectologie traditionnelle (e.g., Wells, 1982), et après une analyse des voyelles du corpus ABI, il est possible de distinguer – certes ceci est très schématique – trois types de réalisation de la voyelle de FACE :

- une monophtongue relativement brève proche du [e] cardinal dans le nord de l'Angleterre (à l'exception de Newcastle) et en Ecosse (Figures 1 et 2),

[2] http://htk.eng.cam.ac.uk/

- une diphtongue fermante dans le sud de l'Angleterre (Figures 3 et 4),
- une diphtongue centripète à Newcastle (Figures 5 et 6).

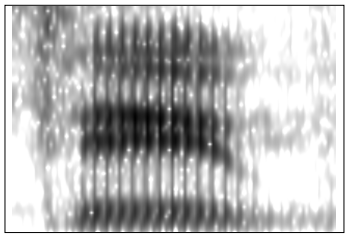

Figure 1 : spectrogramme de la voyelle de *stable* chez un locuteur des Hautes-Terres d'Ecosse. La relative stabilité des formants et l'écart relativement grand entre F1 et F2 indique une monophtongue plutôt fermée.

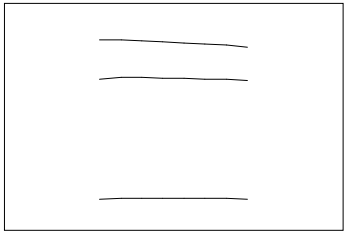

Figure 2 : valeurs centrales des 3 premiers formants de la Figure 1.

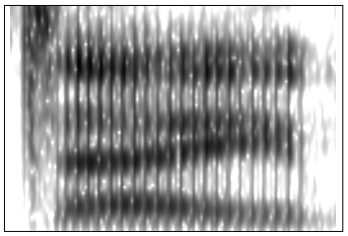

Figure 3 : spectrogramme de la voyelle de *stable* chez un locuteur d'anglais britannique standard. F1 et F2 indiquent clairement une diphtongue fermante.

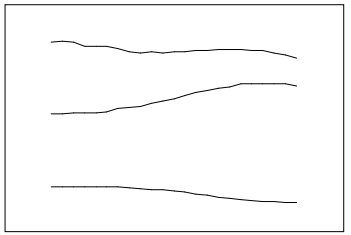

Figure 4 : valeurs centrales des 3 premiers formants de la Figure 3.

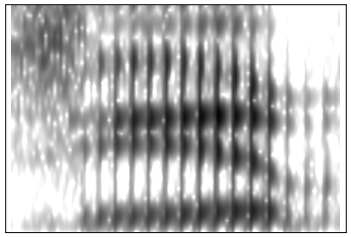

Figure 5 : spectrogramme de la voyelle de *stable* chez un locuteur des de Newcastle. F1 et F2 font apparaître une diphtongue centripète.

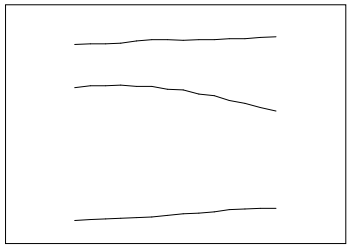

Figure 6 : valeurs centrales des 3 premiers formants de la Figure 5.

Les unités et les échelles ont été délibérément omises sur les figures car il importe de se concentrer sur la structure globale. Nous traitons uniquement du degré de stabilité du timbre au cours de la production de la voyelle, et non pas de la qualité de la voyelle au début et à la fin de son émission. Restait à trouver un moyen économique pour décrire le degré de diphtongaison. Un examen des Figures 1 à 6 appelle deux remarques.

D'abord, la course des formants fait apparaître des courbes plutôt que des droites. Ensuite, quelle information perdrait-on si l'on caractérisait ces (légères) courbes de formants à l'aide de droites ? Cette simplification – ce qu'est toujours un modèle – entraîne-t-elle une perte inacceptable des caractéristiques essentielles de ces voyelles, ou bien permet-elle une représentation économique qui autoriserait en outre l'implémentation d'un système de classification automatique ? Nous avons pris le parti de résumer la trajectoire de chaque formant à une droite par le biais de la régression linéaire. La droite obtenue tend à minimiser la distance entre une valeur formantique effective à un instant donné et le point théorique donné par l'équation de droite. Autrement dit, cette droite est un chemin idéal qui se rapproche autant qu'il le peut des valeurs de formants successives. L'équation obtenue n'est rien d'autre qu'une fonction affine qui renvoie deux termes caractérisant la droite : son ordonnée à l'origine et sa pente. C'est cette dernière valeur que nous utilisons pour caractériser le degré de diphtongaison.

Extraction des paramètres et classification automatique

Les valeurs de F1 et F2 sur toute la durée des voyelles des mots de l'ensemble lexical FACE ont été obtenues automatiquement avec le logiciel Praat[3].

Nous avons réduit notre problème à 14 classes (14 dialectes) à 91 problèmes à 2 classes. Cela se comprend aisément : la voyelle de FACE ne permet pas de distinguer tous les dialectes. Cette réduction du problème permet en outre une modélisation plus rapide et plus maniable. La classification a été réalisée au moyen d'un réseau de neurones artificiels dans son expression la plus simple : le perceptron. Expliquons-en très brièvement le principe général. Le système prend en entrée la valeur de pente de F1 et la valeur de pente de F2, soit deux paramètres ; nous sommes donc dans un espace à deux dimensions. Par ajustement de poids successifs lors de la phase d'apprentissage, l'algorithme tente de trouver la droite dans ce plan qui constitue la meilleure frontière entre les deux classes. Enfin, lors de la phase de classification proprement dite, les valeurs de F1 et de F2 de la voyelle à classer sont présentées au perceptron et, sur la base de ce qu'il a appris, il décide que cette voyelle appartient à la classe A ou B.

[3] http://www.fon.hum.uva.nl/praat/

Résultats et discussion

Il serait rébarbatif de fournir ici une liste exhaustive des taux de classification correcte obtenus. Quelques exemples suffiront à illustrer notre propos. Les taux d'identification sont de 96,83%, 96,25% et 95,19% pour les paires East-Anglia vs. Scottish Highlands, East-Anglia vs. Glasgow et Birmingham vs. Newcastle respectivement. Chacun des membres à l'intérieur d'une paire possède une réalisation très différente ; ces excellents résultats ne sont donc pas surprenants. On peut ensuite les comparer à des taux d'identification non supérieurs au hasard obtenus, par exemple, pour les paires East Yorkshire vs. Lancashire, East-Anglia vs. Standard British English, ou encore Standard British English vs. Inner London (taux avoisinant les 50%). Les mauvais résultats obtenus trouvent un début d'explication dans le fait que, pour ces trois dernières paires, il existe une proximité géographique et/ou historique des dialectes concernés (pour une même paire). Cela ne signifie pas que la voyelle de FACE soit strictement identique, mais cela implique que ce que nous en avons capturé est virtuellement identique pour les deux membres de chacune des paires mentionnées.

Dans cette étude, il apparaît que les valeurs de pentes de F1 et F2 sont suffisantes pour caractériser le degré de diphtongaison de la voyelle de FACE. Plus précisément, il semble que les pentes des deux premiers formants vocaliques suffisent à caractériser la variation inter-dialectale due aux différences de degré de diphtongaison de la voyelle de FACE dans nos données. Notons également que malgré l'apparente parcimonie de la représentation proposée, il est probable que notre modèle soit encore trop redondant : en effet, F1 et F2 sont négativement corrélés, si bien que lorsque les valeurs de l'un augmentent, il est théoriquement possible de prédire que celles de l'autre décroissent, et dans quelle proportion.

Quitte à multiplier les redites, rappelons encore que notre étude vise uniquement à caractériser le *degré de diphtongaison* de la voyelle de FACE, et que nous avons sciemment omis l'information relative aux valeurs absolues des timbres de départ et d'arrivée de la voyelle, timbres qui, comme le notent les ouvrages de références en dialectologie (e.g Wells, 1982), participent à la variation dialectale.

Cette étude nous apprend que la voyelle de FACE est un trait discriminant très robuste dans les dialectes des Iles Britanniques. Cependant, cette approche unidimensionnelle[4] ne dépasse pas le stade de

[4] Certes, nous considérons et F1 et F2, mais la voyelle de FACE ne représente qu'une seule variable dialectologique.

l'exercice de style puisque notre but final est d'intégrer toutes les dimensions (les variables phonético-phonologiques) nécessaires à la résolution du problème à 14 classes. En ce qui concerne l'identification par la machine des dialectes des Iles Britanniques, Barry *et al.* (1989) parviennent à classer 58 locuteurs en 4 aires dialectales grossières avec un taux de classification correcte aux alentours de 74% en utilisant des valeurs de formants. Huckvale (2004) atteint les 71,9% sur le corpus ABI à partir des formants, tous sexes confondus. En employant des paramètres plus élaborés, ses taux de classification augmentent jusqu'à 87,2%. En améliorant la méthode de ce dernier, nous avons obtenus (Ferragne & Pellegrino, 2006) des taux avoisinant les 92%.

Conclusion

La voyelle de FACE est un trait diagnostique particulièrement utile dans une optique d'identification des dialectes britanniques. Les valeurs des pentes de F1 et F2, et l'utilisation d'un perceptron permettent d'atteindre, pour certaines des 91 paires de dialectes étudiées, des taux de classification proches de 100%. Cette étude illustre l'intégration de la connaissance phonétique et dialectologique dans un système d'identification des dialectes par la machine.

Bibliographie

Barry, W.J, Hoequist C.E & Nolan F.J. (1989). 'An approach to the problem of regional accent in automatic speech recognition'. *Computer Speech and Language*, 3: 355-366.

Ferragne, E. & Pellegrino, F. (2006). 'Les systèmes vocaliques des dialectes de l'anglais britannique'. 26^{e} Journées d'Etude sur la Parole, Dinard, 411-414.

Foulkes, P. & Docherty, G. (1999) *Urban Voices. Accent Studies in the British Isles*. Londres : Arnold.

Huckvale, M. (2004). 'ACCDIST: a metric for comparing speakers' accents'. *8th ICSLP*, Jéju, Corée, 29-32

Kortmann, B. & Schneider, E.W. (2004) *A Handbook of Varieties of English*. Berlin : Mouton de Gruyter.

Wells, J.C. *Accents of English.* (1982). Cambridge, UK : Cambridge University Press.