De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux

Martine Adda-Decker

LIMSI-CNRS, Université de Paris-Sud, BP 133, 91403 Orsay CÉDEX, France
Mél : Martine.Adda@limsi.fr

ABSTRACT

This contribution aims at giving an overview of present automatic speech recognition in French highlighting typical transcription problems for this language. Explanations for errors can be partially obtained by examining the acoustics of the speech data. Such investigations however do not only inform about system and/or speech modeling limitations, but they also contribute to discover, describe and quantify specificities of spoken language as opposed to written language and speakers' speech performance. To automatically transcribe different speech genres (e.g. broadcast news *vs* conversations) specific acoustic corpora are used for training, suggesting that frequency of occurrence and acoustic realisations of phonemes vary significantly across genres. Some examples of corpus studies are presented describing phoneme frequencies and segment durations on different corpora. In the future large-scale corpus studies may contribute to increase our knowledge of spoken language as well as the performance of automatic processing.

1. INTRODUCTION

Contrairement à bien d'autres domaines de recherche autour de la parole, la reconnaissance automatique, qui s'effectue sur un flux acoustique continu, nécessite une modélisation de l'ensemble des phénomènes observés dans le signal : au-delà des mots auxquels est associée une représentation de type phonologique dans le dictionnaire de prononciation, il faut modéliser des respirations, des hésitations, des fragments de mots, des brouillons de parole peu ou pas articulés... Dans cette contribution nous allons faire d'abord un rapide état de l'art des systèmes de transcription automatique, présenter leurs performances et analyser les types d'erreurs les plus représentatifs. Nous allons ensuite poser la question de ce que peuvent nous apprendre ces erreurs de transcription. Ceci nous amène à utiliser progressivement les systèmes de transcription comme des instruments d'analyse de grands corpus oraux, d'abord sur les données utilisées pour l'apprentissage ou l'évaluation des systèmes, mais également sur des corpus à visée plus linguistique comme le corpus PFC (Phonologie du Français Contemporain) [15]. De telles études permettent par exemple de décrire et de quantifier des variantes de prononciations [3], des disfluences [9] et des réalisations acoustiques des sons [10]. Quelques adaptations méthodologiques des systèmes s'imposent afin de transformer un système de transcription en un instrument d'analyse de corpus oraux.

2. TRANSCRIPTION AUTOMATIQUE DE LA PAROLE

Il est admis que les progrès réalisés dans le domaine de la transcription ont été largement stimulés par les campagnes d'évaluation, lors desquelles les participants évaluent leur système sur un jeu de test commun et bénéficient de données d'apprentissage communes. Une première évaluation de systèmes de reconnaissance automatique multilingue, incluant les langues française, anglaise et allemande, a été mis en place dans le cadre d'un projet européen (LE-SQALE) en 1994 [26, 32]. Cette évaluation a permis de mettre en évidence pour la transcription automatique, des difficultés spécifiques aux différentes langues. Ainsi le français se caractérise par un nombre d'homophones particulièrement élevé en comparaison avec l'anglais et l'allemand [20]. La simple suite de phonèmes /la/ peut se transcrire suivant le contexte par les mots `la, là, l'a, l'as, las` et l'information acoustique ne permet pas de lever l'ambiguïté à elle seule. Une information sur le contexte de la séquence /la/ est nécessaire. Cette information de contexte est apportée dans les systèmes de transcription par les modèles de langage (des probabilités d'observation de mots et de séquences de mots). Ainsi des séquences comme `tu l'as vu`, `tu la retrouves` sont probables, en revanche des séquences comme `tu l'a` ou `tu las` sont fortement improbables.

Depuis une dizaine d'années, deux campagnes d'évaluation des systèmes de reconnaissance automatique en langue française ont eu lieu [13, 19] et permettent d'apprécier les progrès accomplis. La première campagne d'évaluation a été lancée sous l'égide de l'AUPELF-UREF il y un peu plus de dix ans. Il s'agissait ici de transcrire automatiquement des textes lus du journal *Le Monde* par des locuteurs différents et inconnus en utilisant le corpus BREF [25] enregistré au LIMSI autour des années 90. Ces recherches visaient à démontrer la capacité des systèmes à retrouver les mots prononcés à l'oral. Cette campagne a permis de montrer que des taux de mots erronnés proches de 10% pouvaient être obtenus. Il faut cependant garder à l'esprit que la parole lue est certes médialement de l'oral, mais qu'au fond elle reflète parfaitement la langue écrite. Ceci implique en particulier pour les systèmes, une relative facilité d'estimer des modèles de langage réalistes pour la tâche en question à partir de textes de journaux, dans la mesure où le signal acoustique correspond aux textes lus. Concernant les caractéristiques acoustico-phonétiques de BREF, la lecture à haute voix n'est pas une tâche quotidienne et entraîne pour la plupart des locuteurs une articulation plutôt lente et relativement soignée « qui colle à l'écrit ». Une telle prononciation respecte plutôt bien les hypothèses de modélisation acoustique des mots comme séquences de phonèmes qui est faite par les systèmes de transcription.

La deuxième campagne, ESTER (évaluation des systèmes de transcription enrichie d'émissions radiophoniques), financée par le programme interministériel français TECHNOLANGUE et organisée conjointement par l'AFCP, la DGA et ELDA vise un objectif beaucoup plus ambitieux, dans la mesure où il s'agit ici de parole journalistique

d'émissions radiophoniques, de différentes stations de radio. Même si une bonne partie des enregistrements correspondent à de la parole préparée, i.e. produite à partir d'une préparation écrite, elle est « convertie » à l'oral par des présentateurs et des présentatrices professionnels. Un pourcentage non négligeable des émissions correspond également à des interventions d'invités ou d'auditeurs, pour lesquels il y a souvent peu ou pas de préparation écrite. On est donc ici plus proche de la langue orale. Afin de permettre d'estimer des modèles de langage adaptés à l'oral journalistique (par opposition aux journaux écrits), la DGA, en collaboration avec le LIMSI, a entrepris la transcription manuelle de dizaines voire centaines d'heures de journaux radiodiffusés à la fin des années quatre-vingt dix. Dans cette dynamique a été développé le logiciel TRANSCRIBER [5], qui a trouvé un large succès pour la transcription de corpus oraux, bien au-delà de la communauté du traitement automatique de la parole. L'exercice de transcription manuelle pointe sur des problèmes qui se retrouveront également lors de la transcription automatique. Ainsi on peut se rendre compte à l'écoute attentive que bon nombre de mots sont souvent réalisés de manière incomplète. On pourrait être tenté d'écrire `ad'taleur` pour `à tout à l'heure`, tout comme on peut trouver fréquemment des « trucages orthographiques » comme `'ya` pour `il y a`. Nous avons préconisé pour les transcriptions manuelles le même principe de transcription en orthographe normative comme les conventions du GARS [6], avec un minimum d'indications de prononciations. Ce principe permet le mieux de converger vers des transcriptions stables indépendantes du transcripteur et avec un temps de transcription plus faible que si des annotations spécifiques étaient effectuées. Le problème de « trucages » des prononciations est cependant bien réel et nécessite des adaptations au niveau de la modélisation acoustique des mots.

Le passage de la lecture à la parole radiophonique a donc un impact au niveau des prononciations, avec des réalisations qui peuvent s'écarter de manière plus importante de prononciations canoniques [14]. Les mots outils fréquents, dont l'information est largement portée par le contexte sont souvent mal et peu articulés, et sous l'**effet de répétition** même des mots pleins thématiques (p.ex le mot `architecture` /aRʃitɛktyR/) bien prononcé pendant quatre, cinq fois en début d'émission, finit par être raccourci, ne préservant plus que certaines parties du mot, (comme par exemple [aRʃɛktyR]). Par opposition à une tâche de lecture sans auditoire, qui consiste alors à prononcer les mots écrits de manière relativement équitable, les émissions radiophoniques sont destinées à un large public dispersé et distant, et le souci de compréhension prévaut certainement ici à celui d'une simple articulation claire et équitable.

2.1. Quelques résultats de transcription

La reconnaissance de la parole, consiste à déterminer la meilleure suite de mots $\hat{m}$ à partir de l'observation acoustique x. Avec l'approche statistique ce problème repose alors sur la formule de Bayes

$$\hat{m} = \arg\max_m P(m/x) = \arg\max_m p(x/m)P(m)$$

Le décodeur doit mesurer la probabilité de toutes les suites de mots m possibles pour ce signal : $P(m/x)$. Le problème se transforme grâce à la formule de Bayes en une optimisation à deux termes $p(x/m)P(m)$ pour lesquels des modèles peuvent être estimés à partir de grands corpus d'apprentissage. Le premier terme $p(x/m)$ s'évalue grâce à des modèles acoustiques de mots, construits à partir de modèles acoustiques de phones via un dictionnaire de prononciation. Le deuxième terme $P(m)$ donne une estimation de la probabilité a priori de la séquence de mots m grâce aux modèles de langage N-grammes. La table 1 donne des ordres de grandeurs de quelques paramètres caractérisant les systèmes de transcription.

TAB. 1: Evolution des systèmes de transcription automatique du français : style de parole, données d'apprentissage et taux d'erreur de mots

		apprentissage		*%err.*
campagne	*style*	$p(x/m)$	$P(m)$	*mots*
AUPELF	lecture	BREF	*Le Monde*	12%
1996		100h	40M mots	
ESTER	journa-	radio	journaux, web	11%
2005	listique	100h	400M mots	

Pour les deux évaluations, effectuées sur des données de nature assez différente (lecture enregistrée en laboratoire, parole radiophonique grand public) les taux d'erreur sont proches de 10%. Dans certaines conditions et pour certains locuteurs professionnels les taux peuvent descendre autour de 5%. Mais il est actuellement difficile d'approcher des taux aussi faibles pour une population large de locuteurs. Pour les conversations téléphoniques en français, qui correspondent à un vrai genre oral, les taux d'erreur sont facilement supérieurs à 30% [4]. Certes les conditions acoustiques sont moins bonnes et contribuent à augmenter les erreurs, mais les problèmes essentiels pour la parole conversationnelle concernent à la fois l'estimation d'un modèle de langage approprié au genre traité, et les prononciations des mots avec la modélisation acoustique associée. Des problèmes supplémentaires concernent l'établissement d'une transcription de référence dans des zones disfluentes ou simplement mal articulées.

Par la suite nous allons nous limiter surtout à la parole radiophonique, pour laquelle on peut distinguer deux sources aux problèmes de transcription dans le cadre de la modélisation statistique exposé : est-ce que le système arrive à prédire de manière fiable les mots prononcés ? est-ce que les modèles acoustiques des mots reflètent les prononciations effectivement produites par les locuteurs ? Dans ce qui suit nous allons essayer d'analyser des erreurs de transcription en gardant en tête ces deux questions.

2.2. Analyse des erreurs

Le corpus ESTER dev04 [19] contient 10 heures de parole avec environ 94k mots dont un peu moins de 10k entrées lexicales distinctes. Sur ce jeu de données, le système du LIMSI [21] a produit un peu plus de onze mille erreurs dont sept mille substitutions, environ trois mille omissions et un peu moins de mille cinq cents insertions. Le taux de substitutions est un peu plus que le double des taux d'omissions, qui est lui-même le double du taux d'insertions. Dans l'analyse ci-dessous nous allons aborder plus en détail les erreurs sur les mots les plus fréquents, essayer de caractériser les mots bien ou mal reconnus et examiner la dispersion des erreurs dans le flot de parole via l'étude de zones d'erreur.

Est-ce que les mots fréquents sont bien reconnus ?

L'apparition des mots fréquents dans leur contexte est normalement bien apprise par le modèle de langage $P(m)$ à partir du corpus d'apprentissage. On peut donc s'attendre à un faible nombre d'erreurs dues simplement à un manque d'observation a priori. De même les mots fréquents à l'oral sont bien représentés dans les corpus audio et les modèles acoustiques doivent bien les représenter. Si les corpus ayant servi à l'estimation des modèles de langage et des modèles acoustiques représentent le même genre de données que le test, les mots fréquents devraient

être bien reconnus. Le taux d'erreur moyen sur les 20 mots les plus fréquents est de 10,5% qui est certes endessous du taux moyen de 12% d'erreurs obtenus sur les données de développement (11% pour l'évaluation), mais reste proche du taux d'erreur moyen. Comment expliquer les erreurs de reconnaissance des mots fréquents ?

La table 2 donne la liste des 20 mots (entrées lexicales) les plus fréquents dans le corpus de développement d'Ester. A gauche est donné le nombre d'occurrence de ces 20 mots, classés par rang de fréquence. A droite les mêmes 20 mots sont triés par taux d'erreur intra-classe décroissant (calculé pour chaque mot m_i comme le ratio des mots m_i mal reconnus, incluant substitutions, omissions et insertions de m_i par le nombre de mots m_i dans le corpus de référence). Le taux d'erreur entre parenthèses ne tient compte que des erreurs de substitution et d'omission. On peut voir clairement qu'il y a des tendances très différentes pour ces 20 mots les plus fréquents, qui expliquent à eux seuls plus d'un quart des erreurs commises (28%). Dans le tableau les taux d'erreur les plus élevés correspondent à des mots monophones, donc très courts où le modèle acoustique ne peut pas jouer un rôle discriminant important, et qui admettent en plus des homophones. L'homophonie implique que le choix du mot incombe au modèle de langage (dans l'hypothèse où le modèle acoustique a réussi). Ceci est le cas par exemple pour les paires (`et, est`) et (`à, a`) qui admettent des taux d'erreur autour de 20%. Mais à l'intérieur de chaque paire on peut observer une dissymétrie : `et` et `a` sont beaucoup plus facilement insérés, leur taux d'insertion correspond respectivement à 7,7 et 9,2%. Ceci s'explique par le modèle de langage qui pénalise plutôt l'insertion de `est` et de `à` face à leur contrepartie homophone. Le mot `il`, dont la prononciation canonique dans le système est /il/, a un taux d'erreur élevé de 18,8%. Or `il` est fréquemment réduit à [i] et admet ainsi un quasi-homophone `qui` au rang 19 et au moins deux homophones au-delà du rang 20 : `ils` et `y`, à des rangs supérieurs à 100. Le moins d'erreurs sont observées pour des formes plus longues (2 à 3 phonèmes) qui sont acoustiquement moins ambiguës. Les deux mots les plus fréquents `de` et `la` ont des taux d'erreur faibles de 6,7% et de 3,4%.

TAB. 2: Liste des 20 formes lexicales les plus fréquentes triées par leur nombre d'occurrences et triées par leur taux d'erreur intra-classe. Ce taux d'erreur tient compte des substitutions, omissions et insertions. Le chiffre entre parenthèses néglige les insertions.

forme	*#occ*	*rang*	*forme*	*%err (-%ins)*	*rang*
de	5355	1	et	25,4 (17,7)	4
la	2684	2	est	20,0 (17,1)	8
le	3011	3	a	19,5 (10,3)	14
et	1927	4	il	18,8 (16,2)	15
à	1887	5	à	15,6 (10,2)	5
l'	1840	6	un	13,1 (9,6)	11
les	1800	7	que	9,8 (7,6)	16
est	1367	8	qui	9,6 (7,0)	19
des	1378	9	en	9,6 (7,3)	10
en	1315	10	l'	9,5 (8,3)	6
un	1311	11	les	9,0 (8,3)	7
d'	1116	12	le	8,7 (6,2)	3
du	1101	13	des	8,5 (7,4)	9
a	1815	14	d'	7,8 (6,5)	12
il	916	15	de	6,7 (3,9)	1
que	913	16	une	5,8 (4,3)	18
pour	882	17	dans	5,0 (4,6)	20
une	790	18	pour	4,4 (2,2)	17
qui	797	19	du	4,3 (3,6)	13
dans	724	20	la	3,4 (2,4)	2

En résumé pour les mots fréquents les taux d'erreur au-delà du taux d'erreur moyen s'expliquent essentiellement par deux facteurs : formes courtes et homophonie. Pour les couples de mot (`et, est`) et (`à, a`) particulièrement problématiques, il peut être intéressant au niveau des paramètres acoustiques du système, d'introduire une information prosodique (en particulier l'évolution de la fréquence fondamentale) contenant éventuellement quelques marques distinctives, pour le moment négligées, ainsi qu'un post-traitement morpho-syntaxique spécifique afin de faire progresser les performances. De manière plus générale informations prosodiques et morpho-syntaxiques devraient contribuer à la précision des systèmes de transcription dans le futur.

Quels mots sont les mieux/les moins bien reconnus ?

En examinant les erreurs par mot on se rend compte que la formule

$$\%err = \frac{(sub(m_i^r) + del(m_i^r) + ins(m_i)) * 100}{occ(m_i^r)} \quad (1)$$

n'inclut pas dans sa mesure si le mot m_i est facilement substitué à la place d'un autre mot m_j. Si on veut examiner les causes d'erreur, il peut être important d'inclure dans la mesure les deux types de substitution : le premier type est celui de la formule classique ci-dessus, où le mot m_i de la référence est substitué par le mot m_j de l'hypothèse. Le deuxième type de substitution correspond à la situation inverse où m_j de la référence est faussement transcrit comme m_i dans l'hypothèse. On peut définir une nouvelle mesure de substitutions :

$$sub'(m_i^r) = \frac{\sum_{j \neq i} sub(m_j^h, m_i^r) + \sum_{j \neq i} sub(m_i^h, m_j^r)}{2} \quad (2)$$

ce qui donne la formule suivante :

$$\%err' = \frac{(sub'(m_i^r) + del(m_i^r) + ins(m_i))}{occ(m_i^r)} \quad (3)$$

Le tableau 3 montre des exemples de mots pour lesquels le taux d'erreur augmente beaucoup en passant de la formule classique (*%err*) à la formule modifiée (*%err'*). Ces mots se retrouvent ainsi souvent faussement dans l'hypothèse. La première ligne montre qu'il n'y a pas d'erreur de type *%err* pour le mot `membres` sur les 39 occurrences dans le corpus de référence. Toutes les erreurs concernant le mot `membres` se produisent dans l'hypothèse et sont des erreurs de substitution pour le mot `membre` dans la référence : les probabilités du modèle de langage favorisent la forme au pluriel. Dans un grand nombre de cas le taux *%err'* plus élevé s'explique par des homophones morphosyntaxiques dans la référence, dont les probabilités d'émissions sont plus faibles dans le modèle de langage $P(m)$. Pour le mot `eh` l'explication est une incohérence au niveau des conventions de transcription (`et bien` *vs* `eh bien`) et pour les mots `oui` et `non` pour lesquels les deux taux sont très élevés, les erreurs sont majoritairement des insertions et omissions. Il reste encore des progrès à accomplir pour assurer des taux d'erreur faibles pour des mots importants comme `oui` et `non`. Ces derniers, typiques d'interactions orales, sont relativement peu représentés dans les corpus d'apprentissage. Il est intéressant de voir que `Paris` peut apparaître facilement à la place du mot de référence `pays`. Cette confusion est certainement imputable au modèle de langage, une explication au niveau des modèles acoustiques suggérerait que les mots `Paris, parisien...` peuvent se prononcer de manière proche de `pays`.

Il y a beaucoup plus de formes lexicales pour lesquelles la situation est l'inverse avec un taux %err plus élevé que le taux %err'. En effet tous les mots plutôt rares dans le corpus d'apprentissage, qui admettent un (quasi)-homophone

TAB. 3: Exemples de mots avec *%err < %err'*. Ces mots se retrouvent de manière erronnée dans l'hypothèse. La dernière colonne indique si les erreurs concernent plutôt omissions (del), insertions (ins) ou substitutions (exemples de mots).

forme	*%err*	*%err'*	*#occ.*	*comment.*
Paris	2.8	7.7	71	pays
membres	0	9.0	39	membre
jour	2.6	10.5	38	jours, jouera
reste	2.8	11.1	36	restent
rencontre	2.9	8.6	35	rencontr-es,ent,er
pourrait	3.8	13.5	26	pourraient
cette	3.9	10.2	230	sept, cet, ces
était	16.4	20.8	158	étaient, été, est
non	25.5	35.1	47	del ont
eh	35.3	42.6	34	et bien → eh bien
oui	42.4	48.3	59	del/ins

fréquent ne seront pas facilement proposés à tort par le système de transcription automatique. La table 4 donne quelques exemples.

TAB. 4: Exemples de mots avec *%err > %err'*. Ces mots sont facilement substitués lors du décodage et n'apparaissent que rarement de manière erronnée dans l'hypothèse. La dernière colonne indique si les erreurs concernent plutôt omissions (del), insertions (ins) ou substitutions (exemples de mots).

forme	*%err*	*%err'*	*#occ.*	*comment.*
George	13.0	6.5	23	Georges
Abbas	41.0	21.8	39	Abbass, baisse
Al	51.0	33.0	47	del a, à, Alma
responsables	13.0	6.5	23	responsable
officielle	15.4	7.7	39	officiel
mettre	16.1	8.1	31	mais, promet
cour	18.6	11.6	43	cours, recours
ceux	22.5	15.4	40	del ce
êtes	24.1	15.5	29	est
eu	30.2	21.4	63	eus, vu
ai	40.7	32.7	55	del est
ils	42.8	31.1	201	il, qui
elles	56.7	38.3	30	elle
bon	37.2	26.7	43	mon, ben
me	53.6	46.4	28	del
hein	96.4	76.8	28	del

On peut remarquer que les taux d'erreur sont particulièrement élevés dans la table 4. En utilisant la mesure classique %err on trouve environ 2500 entrées lexicales avec un taux d'erreur de mot supérieur à 25% (8% du corpus et 40% des erreurs). Il s'agit ici, comme nous l'avons déjà vu ci-dessus, soit de mots outils admettant des (quasi)-homophones à fréquence plus élevée. Dans les mots à taux d'erreur très élevés on trouve des mots spécifiques aux discussions orales, comme oui, écoutez, savez, quoi, ben, ah, là, j', ai, suis, , , des mots pleins courts et donc acoustiquement difficile à discriminer comme gens, air, eau, or, sports.

Pour finir sur une note plus positive on peut trouver presque 6000 entrées lexicales parfaitement reconnues (taux d'erreur=0%). Parmi ces mots qui représentent 20% du corpus, on peut trouver aussi bien des mots outils, des mots pleins et des noms propres, comme aujourd'hui, toujours, lors, selon, plusieurs, notamment, soixante, gouvernement, syndicats, secrétaire, membres, coopération, national, Jean, Washington, Bagdad, Pakistan, Rabat, ONU...

On peut remarquer qu'un mot court et acoustiquement difficile comme lors est parfaitement reconnu. Il s'agit ici d'un mot fréquemment utilisé dans le style journalistique et donc favorisé par le modèle de langage. Ceci se fait alors au détriment d'homophones moins fréquent comme le montre l'exemple suivant [1] :

```
REF : sur la tombe d' andré breton il est écrit
je cherche L' OR du temps
HYP : sur la tombe d' andré breton il est écrit
je cherche ** LORS du temps
```

Le pourcentage d'erreurs dû aux homophones morphosyntaxiques (comme membres, membre) est moins important lors de l'évaluation ESTER en 2004 (inférieur à 20%) que pour celle d'AUPELF en 1996 (autour de 30%). Alors que des modèles de langage incluant une information morphosyntaxique n'ont permis d'améliorer le taux d'erreur que de 0,1% (absolu), le fait d'utiliser des modèles de langage incluant un plus grand nombre de trigrammes diminue naturellement les erreurs. La décision du choix entre différents homophones peut se faire plus souvent en tenant compte du contexte plutôt que de faire appel au mécanisme de repli éliminant l'information contextuelle.

Est-ce-que les erreurs se produisent de manière isolée ou en groupe ?

Alors que dans la partie précedente, l'analyse portait sur les mots pris de manière isolée, nous allons examiner ici les erreurs telles qu'elles se produisent dans le flot de parole continue. On peut se poser la question si les erreurs arrivent plutôt de façon isolée ou si une erreur en entraîne d'autres dans son voisinage immédiat, dans la mesure où une erreur risque d'engager le modèle de langage sur une fausse piste pour ses prédictions. Pour éclairer cette question nous avons compté, en plus du nombre d'erreurs, le nombre de zones erronnées dans les transcriptions automatiques du corpus de développement de ESTER 2004, une zone étant définie comme une suite d'erreurs consécutives, les erreurs pouvant être de trois types : substitution, insertion et omission. La table 5 donne un exemple de zone d'erreur de longueur 5. Pour ces dernières il y a 131 de telles zones, qui contribuent avec 655 erreurs à 6% du taux d'erreur.

TAB. 5: Exemple de zone d'erreur de longueur 5. S : substitution, O : omission

REF :	l'	AGGRAVE	ET	PEUT	LE	TUER
HYP :	l'	AGGRAVER		PAUL		TUÉS
ERR :	-	S	O	S	O	S
comm. :		homophone		/pølə/ → [pɔl]		hom.

La figure 1 donne la distribution des zones d'erreur et le nombre d'erreurs par zone en fonction de la longueur des zones. Cette distribution suit en gros une loi de Zipf avec un très grand nombre de zones erronnées de longueur 1 et très peu de zones de longueur élevée. La courbe montre que dans 4000 cas la zone est de longueur 1 et n'entraîne donc pas de dommages collatéraux. Nous avons examiné les types d'erreurs (substitutions, omissions, insertions) en fonction de la longueur de la zone. Il y a environ 65% de substitutions, un peu plus de 20% d'omissions et un peu plus de 10% d'insertions avec des taux de substitutions plus élevés pour les zones de faible longueur et plus faibles pour les zones de longueur élevée. Dans ce dernier cas les taux d'omissions peuvent devenir particulièrement importants.

[1] Alors que le système produit une transcription respectant la casse, la mesure des erreurs est insensible à la casse : tous les mots bien reconnus sont en minuscules, les substitutions sont en majuscules et les insertions/omissions sont marquées par des étoiles. REF, HYP indiquent la transcription manuelle de référence ainsi que la meilleure hypothèse de

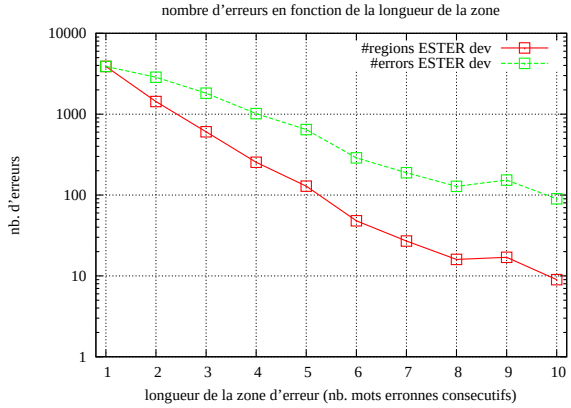

FIG. 1: Nombre d'erreurs et de zones d'erreur en fonction de la longueur de la zone sur le corpus de développement d'ESTER 2004. La longueur d'une zone d'erreur est définie comme le nombre maximum d'erreurs consecutives (incluant substitution, insertion et omission).

L'exemple suivant, avec 3 zones d'erreurs (2 zones de longueur 1 et 1 de longueur 4) permet d'illustrer quelques cas-types d'erreurs.

```
REF : ELLE atteindront sur le bassin parisien
PRÈS DE LA LOIRE    sur le sud de la bretagne la
vendée vingt trois À vingt quatre degrés
HYP : ELLES atteindront sur le bassin parisien
** ** ** PRÉVALOIR  sur le sud de la bretagne la
vendée vingt trois * vingt quatre degrés
```

La première zone d'erreur de longueur 1 correspond en fait à une erreur de la transcription de référence et le système rétablit ici une erreur dans la référence [2]. On peut estimer autour de 2% le taux d'erreur résiduel humain. Ce taux varie évidemment en fonction du soin et du coût dépensés pour la production de la transcription de référence. La zone d'erreur de longueur quatre implique un syntagme prépositionnel `près de la Loire`, dont le spectrogramme est montré dans la figure 2. L'hypothèse `prévaloir` à trois syllabes peut être considérée comme spectralement et surtout temporellement acceptable, étant donnée la réalisation du syntagme en question sous forme de trois syllabes sous une forme comme [pRɛ.dla.lwaR] ou [pRɛ.la.lwaR]. La suite de mots outils `près de la` est articulée rapidement en 300 ms (environ 50 ms pour le mot `de`), la même durée que le noyau du syntagme `Loire`. La dernière zone d'erreur correspond à une élision du mot outil (monophonème) `à` dans un contexte vocalique gauche identique (`trois /tRwa/`). La durée mesurée sur la séquence /aa/ enchaînée est inférieure à 100 ms. Ceci est une situation particulièrement favorable aux élisions si les deux voyelles sont enchaînées sans césure [3].

Nous terminerons cette partie par une sélection d'extraits montrant différentes situations d'erreurs. Ces exemples illustrent que les décisions erronnées du système n'ont en général pas d'explication simple, dans la mesure où la décision est prise en fonction de la réalisation du locuteur en appliquant conjointement modèles acoustiques et modèle de langage.

Dans l'exemple suivant on trouve le même type d'erreur de suppression de voyelle que précédemment, dans la séquence `déjà à`. La liaison entre `elles` et `appellent`, bien que autorisée par le système grâce à

transcription automatique respectivement.

[2]La forme `elles` arrive à être reconnu face à son homophone plus fréquent `elle`, car le mot suivant `atteindront` marque acoustiquement le pluriel et le trigramme `elles atteindront sur` existe dans le modèle de langage.

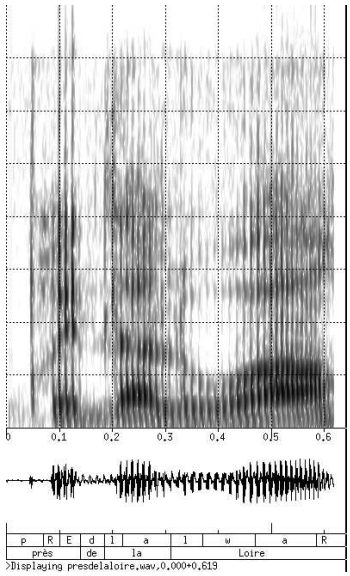

FIG. 2: Spectrogramme de la suite `près de la Loire` reconnue comme `prévaloir`. On observe une chute presque complète du mot `de`, esquissé uniquement par l'occlusion voisée du /d/, la phase d'explosion correspond au /l/ de `la`.

un phonème /z/ optionnel, provoque une zone d'erreur de longueur 3 qu'on doit attribuer plutôt au modèle de langage (cf. erreur `elles, elle` dans la table 4) qu'à un problème de modélisation acoustique.

```
REF : ELLES ** APPELLENT déjà À DES MOBILISATIONS
dans le public
HYP : ELLE S' APPELLE déjà * *** DÉMOBILISATION
dans le public
```

L'exemple suivant illustre la différence de précision de transcription automatique que l'on peut observer sur des mots qui ponctuent l'oral (ligateurs, incises...) : `alors, hein, je le signale`. On peut remarquer que l'hypothèse produite est significativement plus courte, ce qui reflète à l'écrit la réalisation acoustique contractée, réduite de ces mots ou suites de mot.

```
REF : ALORS françois chérèque HEIN JE le SIGNALE
qui sera l' invité de
HYP : À françois chérèque **** ** le SIGNAL qui
sera l' invité de
```

Comme déjà évoqué dans l'introduction, l'effet de répétition entraîne de la part des locuteurs, des performances de prononciation éloignée des prononciations standard, canoniques. Ci-après une réalisation contractée de `ministre de`, reconnue comme `minimiser`. Cette erreur est clairement d'origine acoustique, le modèle de langage favorisant sans ambiguïté le `ministre de l'intérieur`.

```
REF : nicolas sarkozy MINISTRE DE l' intérieur
bonjour
HYP : nicolas sarkozy *** MINIMISER l' intérieur
bonjour
```

Ici encore une erreur d'origine plutôt acoustique : la contraction de `c'est les` en `ses` dans l'hypothèse met en évidence une articulation rapide et peu précise sur ces suites de mots outils.

```
REF : pourtant et C' EST LE paradoxe les nouveaux
moyens
HYP : pourtant et * ** SES paradoxe les nouveaux
moyens
```

L'analyse des erreurs de transcription montre que bon nombre d'erreurs sont simplement dus aux prédictions

insuffisantes des modèles de langage. En particulier dès lors que plusieurs homophones sont en compétition, le choix ne repose plus que sur le modèle de langage. En revanche sur les erreurs où des prononciations différentes sont proposées entre référence et hypothèse, la modélisation acoustique influe sur la décision. Examinant différents types de parole (présentations journalistiques ici, à comparer d'un côté avec la lecture et d'un autre côté avec des conversations) et en écoutant le signal sur les parties erronnées comme sur des parties bien transcrites, il apparaît que les prononciations varient beaucoup en fonction du style de parole. Ces variations n'impliquent pas nécessairement des erreurs. En effet les modèles acoustiques de phones en contexte intègrent de manière implicite une grande partie des variantes [1]. En passant de la lecture à des conversations privées il est cependant nécessaire d'ajouter davantage de variantes dans le dictionnaire de prononciation. Il s'agit essentiellement de l'adjonction de formes contractées par rapport à la forme canonique (pleine) pour les mots outils très fréquents, pour les nombres, les dates et autres locutions. De nombreuses erreurs de transcription impliquent encore le schwa final comme dans `de, le, je, ne, se, ce, que, te, me`, dont la chute n'est pas prévu pour l'instant devant consonne. Or dans ces mots CV monosyllabiques à noyau faible le schwa tombe très facilement dès lors que le mot précédant se termine par une voyelle. Ainsi dans la séquence `près de la` le schwa du `de` est fréquemment omis à cause de la voyelle finale /ɛ/ du mot `près` qui précède. D'autres voyelles, consonnes, clusters voire des syllabes entières peuvent disparaître ou les syllabes peuvent se restructurer. Il est connu que les liquides sont facilement élidées ainsi que la consonne /v/. Ces observations amènent naturellement à se poser des questions sur les régularités d'apparition de ces variantes de prononciations et de manière plus générale sur les spécificités de l'oral par rapport à l'écrit oralisé, l'usage des mots, des tournures de phrases, des hésitations et autres disfluences. Il est important d'essayer de tirer profit des outils de transcription automatique et des grands corpus accumulés pour l'apprentissage des modèles du système afin d'en dégager des nouvelles connaissances sur l'oral.

3. ANALYSE DE CORPUS

Nous abordons ici l'analyse de corpus oraux en gardant parmi les objectifs de mieux cerner ce qui peut poser problème aux systèmes de transcription automatique. Nous avons souligné le problème de réalisations acoustiques réduites et contractées, provoquant des erreurs de transcription, et nous avons évoqué le besoin d'introduire des variantes de prononciations en fonction du style de parole. Ces problématiques sont convergentes à des objectifs de recherche en phonétique et phonologie, comme l'étude de la variation phonologique en français, en particulier la liaison et le schwa [12, 17, 16]. Des premiers travaux quantitatifs et qualitatifs sur la réalisation du schwa et de la liaison [8] ont déjà été effectués sur grands corpus, travaux qui viennent compléter des études plus fines sur corpus contrôlés [18, 28]. Plus récemment nous nous sommes intéressés aux restructurations syllabiques en parole spontanée faiblement contrôlée, aux disfluences dans le même type de corpus, aux hésitations en différentes langues [30, 10]. Des études des triangles vocaliques à partir de grands corpus ont montré l'impact de la durée sur la réalisation des voyelles, montrant un mouvement centripète des valeurs formantiques pour des durées de segment décroissantes. Ce résultat est à mettre en relation avec la discussion des erreurs de transcription présentée dans la section précédente et l'étude de la durée des segments ci-dessous. Cette tendance semble être indépendante de la langue [22]. Des études des variantes de prononciations en fonction de l'accent régional sont en cours dans la communauté autour du projet PFC (Phonologie du Français Contemporain) [15] avec un corpus visant à réunir quelques centaines d'heures de variétés de français en différents styles.

Dans ce qui suit nous allons utiliser différents corpus pour aborder la question de la fréquence des phonèmes pour la langue française [11, 27, 7] sur de grands corpus, et vérifier s'il y a des variations en fonction du genre de corpus. Des distributions contextuelles seront ensuite présentées. En effet pour la transcription automatique les fréquences des phonèmes, et en particulier les fréquences de phonèmes en contexte sont exploitées pour la modélisation acoustique. Ainsi le passage de modèles acoustiques de phones hors contexte à des modèles triphones (tenant compte des phonèmes gauche et droit) entraîne un gain significatif en précision de transcription. Nous présentons ensuite une étude comparative des durées segmentales entre genres de corpus et, à la fin une étude prosodique qui malgré son statut préliminaire permet d'envisager un large éventail de travaux à base de corpus étiquetés.

3.1. Corpus utilisés

Dans les études présentées ci-après nous faisons appel à trois corpus : un corpus de parole journalistique de 25 heures provenant de différentes stations de radios (*France Inter, France Infos*) et de chaînes de télévision (*France2, France3*), un corpus de parole conversationnelle par téléphone (entre amis, membres d'une même famille, ainsi qu'entre inconnus) et une partie du corpus PFC (phonologie du français contemporain), incluant différents styles de parole, avec de la lecture (liste de mots et texte) et des entretiens entre connaissances. Pour PFC nous considérons 12 points d'enquête, notamment Aveyron-Paris [3], Biarritz, Brécey, Brunoy, Dijon, Douzens, Lacaune, Lyon-Villeurbanne, Nyon, Roanne, Rodez et Vendée. La parole journalistique est caractérisée par le fait qu'il s'agit de parole préparée et publique, s'adressant à un public hétérogène n'interagissant pas (ou très peu) avec le locuteur. Le locuteur a souvent le monopole de la parole. Pour les conversations téléphoniques, il s'agit de dialogues privés où les interlocuteurs négocient leur tour de parole, mais ne disposent que du canal audio pour la communication. La situation téléphonique très naturelle fait que les interlocuteurs oublient facilement que la parole est enregistrée. Le corpus PFC contient de la parole lue et des conversations entre deux ou plusieurs personnes d'un même cercle de connaissances autour d'un micro. La durée et des facteurs caractérisant la production de l'oral sont donnés dans la table 6 pour les différents corpus.

TAB. 6: Caractéristiques des corpus oraux examinés.

corpus	*durée*	*parole/ interaction*	*auditoire*
radio TV	25h	préparée $\sim$ monologue	large public
convers. tél.	120h	spontané dialogue	1 ami
PFC	32h	lu+spont.	$<= 3$ pers
- mots	6h	lecture monologue	-
- texte	5h	lecture monologue	-
- entretien	21h	spontané dialogue	$<= 3$ pers

[3] Le point d'enquête Aveyron-Paris regroupe des locuteurs aveyronnais vivant depuis de nombreuses années à Paris.

3.2. *Fréquence des phonèmes*

Les corpus ont été alignés phonémiquement par le système de transcription automatique. Le dictionnaire de prononciation ne contient que des prononciations canonique incluant liaisons et schwa optionnels. Les fréquences des phonèmes sont ainsi calculées sur les trois corpus. La figure 3 donne les pourcentages des voyelles dans les corpus. On peut voir que les trois types de corpus suivent globalement la même évolution. Nous avons pris la courbe du corpus journalistique (en rouge) comme référence, pour laquelle nous avons indiqué les pourcentages des voyelles sur l'axe Y. Sur les axes X et Y, le schwa /ə/ et la voyelle centrale ouverte /œ/ sont codés ensemble par le symbole `x` ; la voyelle centrale fermée /ø/ est codée `eu`. Les deux /o/ ouvert et fermé sont comptés ensemble (o,c) et occupent ainsi 3,6% du corpus, chacun représentant environ 1,8%. Pour PFC nous donnons une courbe globale intégrant les divers types de parole de 12 points d'enquête. On peut observer que pour les conversations téléphoniques on a coté voyelles antérieures fermées, 1% (absolu) de /e/ en moins par rapport à la parole journalistique et pour les voyelles ouvertes, environ 2% de /ɛ/ et 2% de /a/ en plus, ainsi qu'un peu moins de 1% de /o,ɔ/ en moins. Pour les voyelles nasales il y a surtout un petit déficit pour la voyelle /ã/ dans les conversations téléphoniques. La courbe PFC reste proche des deux autres courbes et nous allons examiner les fréquences d'occurrences ici en fonction des différents styles (voir figure 5). Les entretiens PFC (guidés et libres) sont comparés aux conversations téléphoniques dans la figure 5 (à gauche). Deux courbes supplémentaires sont rajoutées afin de comparer quelques points du Nord (Brécey, Brunoy, Vendée, Dijon) à deux points du Sud (Douzens, Lacaune). Contrairement aux conversations téléphoniques, les entretiens PFC ont des taux de /e/ et de /ɛ/ sensiblement identiques. Il y a environ 2% de schwa supplémentaires dans les conversations téléphoniques que dans les entretiens PFC ; la courbe PFC-Sud (Douzens et Lacaune) a environ 2% de schwa en plus que la courbe PFC-Nord (4 points d'enquête). Dans la figure 5 à droite nous comparons la courbe PFC globale au sous-ensemble formé par le texte lu. Il est intéressant de noter que la courbe du texte s'écarte significativement des autres courbes par un déficit de /a/ (plus de 2% en absolu). En revanche il y a un nombre important de schwa, qui fait passer légèrement le schwa devant le /a/, le schwa devenant ainsi le phonème le plus fréquent. Ceci est dû à la forte présence de schwas dans le sud, qui est une des marques de l'accent méridional. Il faut cependant garder à l'esprit que le texte PFC *Le maire de Beaulieu* a été construit entre autre pour l'étude du schwa dans le français régional. Cette construction a eu un effet non négligeable sur la distribution des voyelles.

La figure 4 montre le pourcentage des consonnes du français observés dans les trois corpus. Globalement les courbes d'occurrence des consonnes suivent la même évolution sur les différents genres de corpus. Les consonnes les plus fréquentes du français sont /R/, /l/, /s/, /t/. On voit que le classement varie légèrement en fonction des genres de corpus examinés. Concernant la parole spontanée on peut voir que la parole téléphonique génère des proportions de /m/ et de /w/ plus élevées, ce qui est largement dû, pour le /m/ à une proportion élevée de mots comme `mais, moi, me` et des interventions de « backchannel » `hum` particulièrement élevé ici. Le /w/ provient de nombreuses occurrences de `oui, ouais, moi, voilà, toi, vois, crois`. On peut remarquer que ces mêmes mots enrichissent les statistiques du /a/ et du /ɛ/, comme nous avons pu le constater côté voyelles.

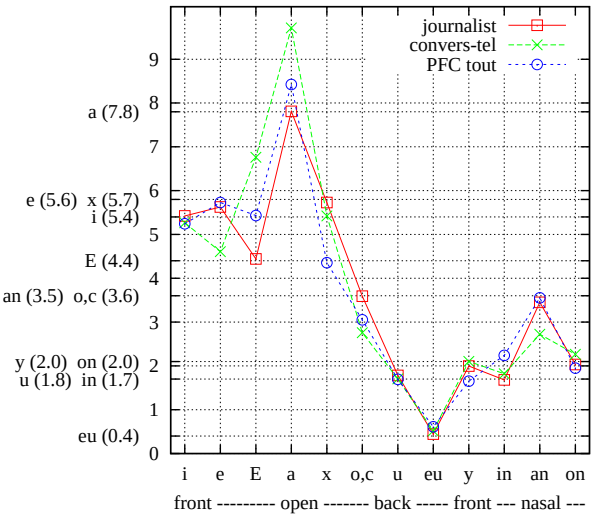

FIG. 3: Pourcentage d'occurrence des voyelles sur 3 types de corpus **journalistique, conversations téléphoniques, PFC divers, PFC**. Sur l'axe Y sont reportées les voyelles avec leur pourcentage du corpus journalistique, établissant ainsi une échelle de classement pour les voyelles.

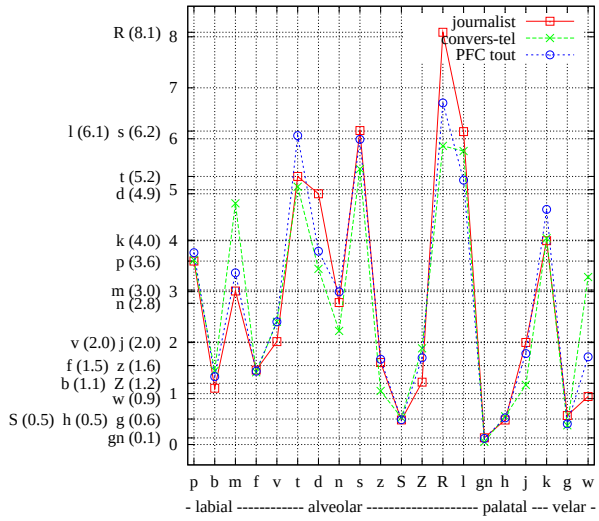

FIG. 4: Pourcentage d'occurrence des consonnes données sous forme de pourcentage de corpus. Sur l'axe Y l'échelle de classement des consonnes suit le corpus journalistique.

3.3. *Fréquence d'occurrence des voyelles en contexte*

Les fréquence d'occurrence des voyelles en contexte ont été mesurées sur les trois types de corpus : journalistique, conversations téléphoniques et PFC. Par manque de place nous ne présenterons ici que les mesures pour le corpus journalistique (voir Figure 10). Concernant les contextes nous avons distingué 6 classes consonantiques mélangeant pour ces classes des critères articulatoires, de fréquence d'occurrence et d'économie de présentation. Nous avons utilisé les lieux d'articulation tels qu'ils sont décrits dans le tableau synthétique des consonnes de l'IPA et nous avons considéré les liquides /l/ et /R/ très fréquentes en français séparément . A chaque classe est associé un code couleur dans les histogrammes : labial (/p/, /b/, /m/, /f/, /v/) en rouge, alvéolaire (/t/, /d/, /n/, /s/, /z/) en vert, postalvéolaire (/ʃ/, /ʒ/) en bleu foncé, liquide /R/ en rose, liquide /l/ en bleu ciel et palato-vélaire (/gn/, /ɳ/, /j/, /k/, /g/, /w/) en jaune. La dernière classe de consonnes palato-vélaires a été motivée surtout par une volonté de limiter le nombre de classes à six, pour des raisons de présentation des résul-

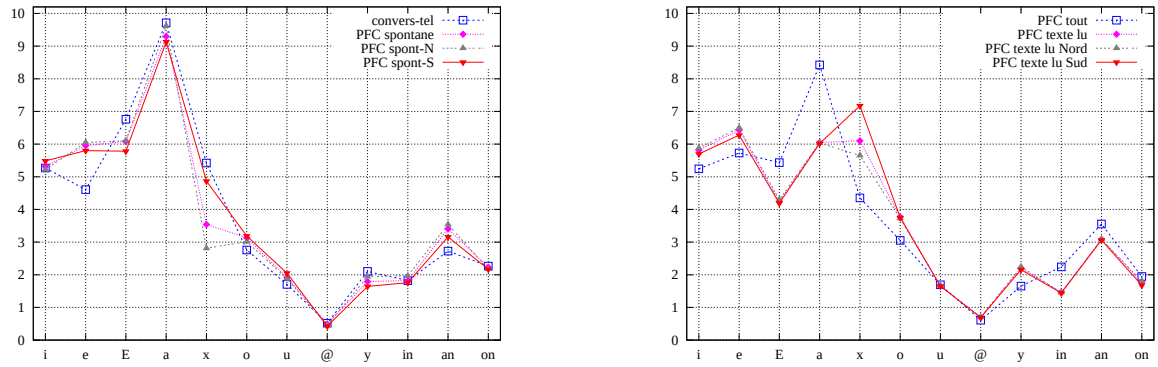

FIG. 5: Fréquences d'occurrence des voyelles. **A gauche :** parole spontanée :conversations téléphoniques et entretiens PFC. **A droite :** corpus PFC (tout et parole lue) : les fréquences d'occurrence des voyelles sont obtenues sur 12 points d'enquête. Pour le texte lu 2 courbes sont rajoutées comparant quelques points du Nord (Brécey, Brunoy, Vendée, Dijon) à deux points du Sud (Douzens, Lacaune).

tats. Ainsi pour un corpus donné les mesures tiennent sur une page sous forme d'un jeu de 6 histogrammes pour les 6 contextes consonantiques gauches. Ensuite dans chaque figure, ayant un contexte gauche fixé, on peut voir pour chaque voyelle la fréquence d'occurrence de chacune des 6 classes de contextes consonantiques droits. Ici les fréquences d'occurrences correspondent aux comptes dans les corpus sans ramener les comptes aux pourcentages d'occurrence.

Sans vouloir trop commenter ces mesures brutes, on peut faire quelques observations. Le contexte gauche alvéolaire (histogramme du milieu à gauche) est le plus riche en occurrences et le plus varié en contextes droits. En particulier on peut observer que cette classe apparaît fréquemment en même temps à gauche et à droite (contributions vertes importantes dans les barres des histogrammes), alors que ceci est beaucoup moins vrai pour les autres classes (par exemple peu de rouge dans la première figure en haut à gauche pour le contexte gauche labial, peu de rose dans la figure en haut à droite pour les contextes gauches de /R/ etc.). Contrairement aux données journalistiques, les conversations téléphoniques ont les fréquences les plus élevées pour le contexte gauche labial et les voyelles ouvertes /a/ et /ɛ/, ce qui rejoint nos observations sur les phonèmes hors contexte. Pour le corpus journalistique, le contexte gauche alvéolaire est le plus productif pour pratiquement toutes les voyelles, mise à part le /u/ plus fourni pour le contexte gauche labial, et la voyelle nasale /ɔ̃/ qui émerge particulièrement dans le contexte gauche palato-vélaire avec les deux consonnes /j/ et /k/ (avec des mots comme `question, millions, région, contre, compte, conseil`). Pour les 3 corpus (journal, conversations, PFC) on note des comptes élevés de la voyelle /a/ en contexte palato-vélaire gauche. Ceci est dû aux nombres comme `trois, quatre, quarante, soixante` et aux mots fréquents comme `quoi, crois, voilà, soir, avoir`. Sur les 18k contextes *palato-vélaire*-/a/ répertoriés pour le corpus journalistique, 8k proviennent de contextes /w/-/a/. Comme pour les fréquences des phonèmes, nous trouvons ici des différences assez marquées entre genres de corpus, variations qui sont à mettre en lien avec les variations du lexique. Des études plus fines sont nécessaires dans le futur pour compléter ces premières mesures présentées ici. Elles illustrent cependant que le rang de fréquence des voyelles varie en fonction du contexte phonémique gauche et droit.

3.4. Durées

Si nous pouvons mesurer des différences importantes de durées sur les segments phonémiques entre corpus de différents styles, ceci suggère qu'il existe bel et bien des différences significatives entre les prononciations qui contribuent à dégrader les performances lors de la transcription automatique. A partir de l'alignement automatique avec des dictionnaires de prononciation standard (incluant peu de variantes) la figure 6 donne la distribution des segments phonémiques en fonction de leur durée pour les corpus journalistique et de conversations téléphoniques. Plus le maximum de la distribution se trouve décalé vers la gauche (i.e. vers les segments courts) plus la courbe indique un risque de désaccord entre prononciations standard attendus et prononciations effectivement réalisées par les locuteurs, ces réalisations présentant alors potentiellement des réductions temporelles. Pour ces dernières on peut chercher des explications, la première articulatoire : le phonème, facile et rapide à réaliser, a une durée intrinsèque courte. Un débit rapide avec une articulation incomplète réduit alors encore cette durée. Une deuxième explication peut venir des fréquences d'occurrence : une observation très fréquente est une observation à contenu d'information faible et risque donc d'être négligée dans le signal acoustique. Il est fort plausible que dans une parole spontanée ces deux facteurs se trouvent combinés.

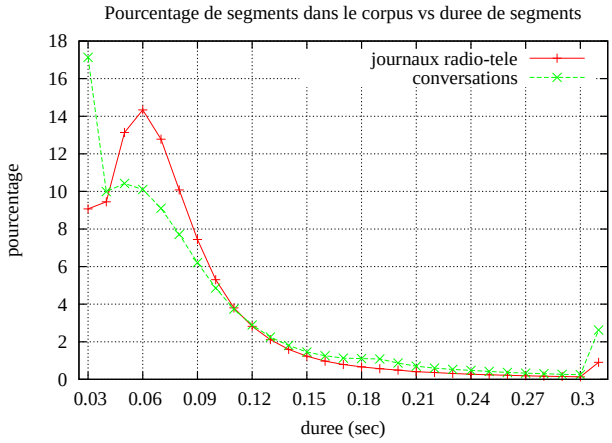

FIG. 6: Distribution de la durée des segments phonémiques pour deux styles de parole : parole préparée d'émissions journalistique et parole spontanée de conversations téléphoniques.

La figure 6 montre clairement l'effet du style de corpus. Pour le corpus journalistique la distribution de durée des segments réalise un pic globalisant 14% du corpus à 60 ms, alors que pour la parole conversationnelle ce pic se trouve à 50 ms avec seulement 10% des données. Mais pour ce dernier le maximum des durées est concentré sur

les durées courtes avec 17% des segments sur la durée minimale de 30 ms imposée par l'alignement automatique (modèles de Markov cachés à trois états avec transition entre états correspondant à 10 ms). Pour la parole journalistique ce taux est « seulement » de 9%. Tous les phonèmes ne contribuent pas tous de manière égale. La table 7 montre les phonèmes réalisant les plus forts pourcentages de durée minimale dans le corpus de conversations téléphoniques. Des taux plus faibles sont observés pour le corpus journalistiques, mais l'ordre des phonèmes impliqués reste sensiblement identique. Concernant les consonnes on trouve en premier lieu le /l/ avec 43% des segments à durée minimale. Le /l/ est très fréquent en français, en particulier sur les mots outils [4]. De manière générale les liquides, les semi-voyelles, le /v/, le /d/ et le /n/ sont les consonnes les plus affectées. Le /d/ comme le /l/ sert fréquemment de support aux mots outils. Pour les voyelles on voit que le schwa arrive en première position avec pratiquement 50% des occurrences. Ceci confirme par mesures détournées la nature instable du schwa. Ce qui est plus intrigant c'est de voir des pourcentages élevés pour les voyelles ouvertes /ɛ/ et /a/. Concernant le /a/ nous pouvons observer par exemple que ce segment peut quasiment disparaître dans des mots ou séquences de mot comme `voilà, parce que` réalisés plutôt comme `v'là, p'ce que`. De même les voyelles antérieures fermées sont alignées fréquemment avec une durée minimale. On peut remarquer que bon nombre de phonèmes qui ont parmi les plus forts pourcentages de durée minimale, sont également parmi les plus fréquents. Il manque cependant les consonnes non-voisées comme /t/ et /s/ qui ont une durée intrinsèque plus longue.

TAB. 7: Phonèmes réalisant les plus forts pourcentages de durée minimale dans le corpus de conversations téléphoniques. Ces pourcentages sont donnés pour les voix d'hommes et les voix de femmes.

cons	*% durées min.*		*voy*	*% durées min.*	
phon	*hom*	*fem*	*phon*	*hom*	*fem*
l	43	44	ə	50	46
ɥ	34	43	ɛ	30	24
v	33	26	a	26	16
j	28	21	i	25	20
R	26	23	y	24	27
d	23	17	e	24	16
n	20	12	ø	22	21

Cette première analyse des durées met en évidence un phénomène de réduction temporelle en parole spontanée et en donne une première quantification. Ce phénomène qui est certes connu, nécessite cependant dans le futur des investigations plus approfondies en lien avec des études sur les variantes phonologiques du français, la prosodie, le débit et la modélisation des prononciations pour le traitement automatique. En particulier des études en fonction des contextes phonémiques et des mots supports permettront d'éclairer si la réduction temporelle est plutôt conditionnée par les contextes phonémiques ou par la fréquence lexicale.

3.5. Prosodie

Nous terminons par un travail engagé dans le contexte du projet CNRS TCAN Varcom et du projet ANR PFC-Cor sur les variétés régionales du français. Le corpus PFC vise à proposer des échantillons représentatifs des parlers normatifs et vernaculaires d'un grand nombre de variétés de l'espace francophone via des enregistrements de parole lue et d'entretiens libres et guidés. Nous proposons ici une étude montrant comment à partir de grands corpus et de traitement automatique on peut partir à la recherche d'indices prosodiques caractérisant les accents régionaux.

L'accent suisse, représenté par le canton de Vaud (autour de la ville de Nyon), a pu être identifié facilement face à des parlers français méridionaux (Aix-Marseille, Douzens, Biarritz) et du nord (Vendée, Brécey) lors de tests perceptifs [31] autant par des sujets français de la région parisienne, que par ceux de la région d'Aix-Marseille. Est-ce que cette facilité d'identification pourrait être liée à des aspects prosodiques ? Une étude de la distribution des durées de segments (pris globalement, mais aussi en distinguant voyelles et consonnes) ne montre pas de différence significative pour le pays de Vaud. Or, d'après différentes études (p.ex. [23]) les schémas intonatifs du canton de Vaud semblent présenter quelques particularités, en particulier une tendance à l'accentuation de la première syllabe d'un mot bisyllabique, qui entraîne en général une montée de la courbe mélodique. D'après ces études cette tendance peut être liée au substrat franco-provencal, caractérisé par une accentuation de la pénultième syllabe. Eventuellement, sur un plan diatopique, ceci peut trahir un contact avec la langue allemande.

Le corpus du texte lu, segmenté à la fois en mots et en phonèmes permet d'étudier l'évolution de la courbe de F0 aux frontières de phonèmes et de mots. Un étiquetage en parties du discours permet d'affiner les analyses. A partir de la segmentation phonémique automatique de douze points du corpus PFC (notamment Aveyron-Paris [5], Biarritz, Brécey, Brunoy, Dijon, Douzens, Lacaune, Lyon-Villeurbanne, Nyon, Roanne, Rodez et Vendée), nous avons pu mesurer à l'aide de PRAAT [29] la fréquence fondamentale moyenne par segment. Dans l'idée d'étudier de manière plus détaillée le bigramme `déterminant nom`, nous avons ensuite associé des parties du discours aux mots du texte lu. La table 8 montre le nombre de bigrammes extraits par région, pour lesquels la fréquence fondamentale est définie à la fois sur la voyelle du déterminant et sur la voyelle de la première syllabe du nom.

TAB. 8: Nombre de bigrammes `déterminant nom` extraits du texte PFC.

Point d'enquête	*#bigrammes Det Nom*
Aveyron-Paris	155
Biarritz	186
Brécey	188
Brunoy	188
Dijon	137
Douzens	194
Lacaune	194
Lyon-Villeurbanne	193
Nyon	233
Roanne	161
Rodez	105
Vendée	93

La figure 7 montre la distribution des séquences *dét-nom* en fonction de la différence de F0 (ΔF_0) mesurée entre la première voyelle du nom et la voyelle du déterminant (qui précède immédiatement). Les mesures sont obtenues à partir du sous-corpus de PFC correspondant au texte lu qui est le même pour tous les locuteurs. Pour cette étude les voix d'hommes seules sont exploitées. La courbe en gras correspond à la distribution moyenne obtenue à partir

[4] Ceci peut être mis en relation avec des erreurs observées dans la première partie (p.ex. la contraction de `c'est les` en `ses`).

[5] Le point d'enquête Aveyron-Paris regroupe des locuteurs aveyronnais vivant depuis de nombreuses années à Paris.

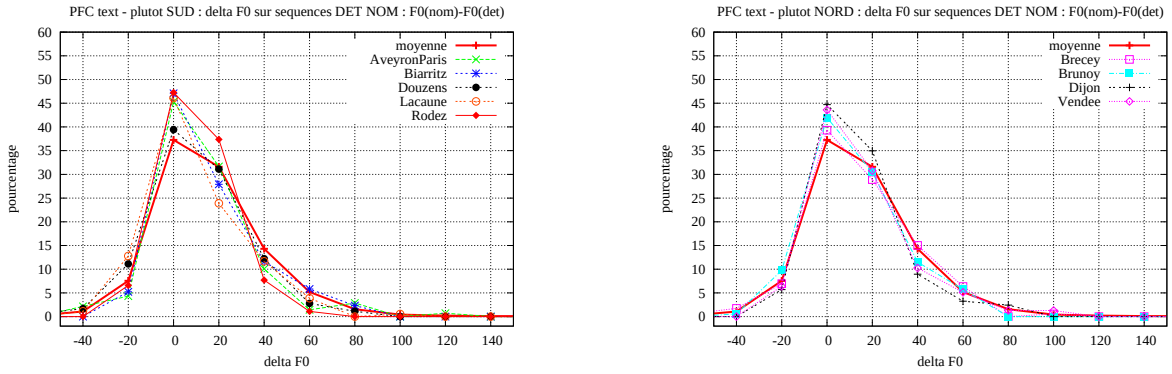

FIG. 7: ΔF_0 sur les séquences de partie de discours **déterminant nom** à partir du texte lu du corpus PFC ; **à gauche :** points d'enquête localisés plutôt au sud (Aveyron-Paris, Biarritz, Douzens, Lacaune, Rodez) ; **à droite :** points d'enquête localisés plutôt au nord (Brécey, Brunoy, Dijon, Vendée).

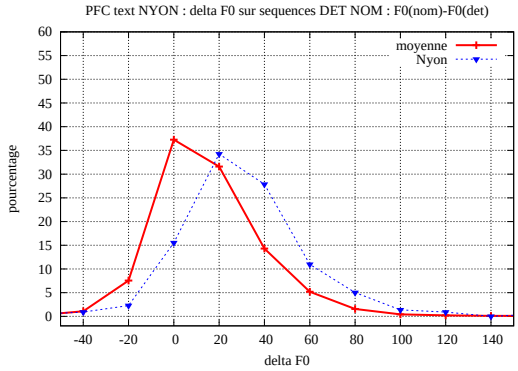

FIG. 8: ΔF_0 sur les séquences de partie de discours **déterminant nom** à partir du texte lu du corpus PFC : courbe moyenne et Nyon (pays de Vaud).

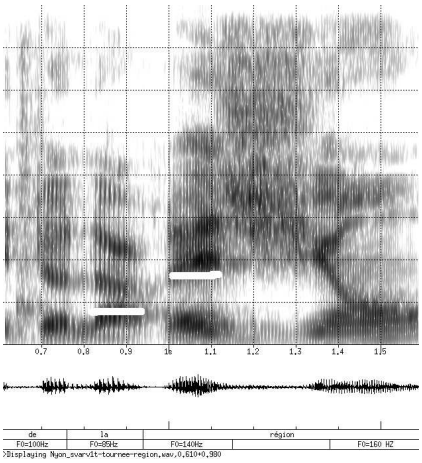

FIG. 9: Spectrogramme illustrant un ΔF_0 positif sur la séquence *dét-nom* : `la région` du locuteur `svarv` du pays de Vaud, extraite du corpus PFC.

de 12 points d'enquête. A gauche on voit les points d'enquête du sud et à droite plutôt du nord de la France. Les courbes ne varient pas significativement en fonction d'une séparation nord/sud. Les courbes montrent que sur les séquences *dét-nom* en français standard, ΔF_0 est centrée autour de 0 et reste proche de 0. Ainsi presque 40% des séquences *dét-nom* ont un ΔF_0 de zéro et pour 30% des données ΔF_0 vaut 20Hz. Seulement 15% des séquences *dét-nom* ont un ΔF_0 de 40Hz. La figure 8 compare la courbe moyenne au canton de Vaud. Le français du pays de Vaud se distingue dans cette étude par une montée de F0 qui commence dès la première syllabe du nom : 35% des 233 séquences examinées ont un ΔF_0 de 20 Hz et presque 30% atteignent 40Hz.

Ce premier résultat semble indiquer une spécificité prosodique qui irait dans le sens d'une accentuation de la première syllabe (ou amorce d'accentuation dès la première syllabe). Des études sur plus de données et avec des patrons plus complexes qu'un simple bigramme de partie de discours sont prévues, en particulier pour l'étude de la pénultième syllabe du nom.

4. CONCLUSION

Dans cette contribution nous avons essayé de donner un aperçu de quelques problématiques de recherche en transcription automatique et en analyse de grands corpus oraux de différents genres, avec des perspectives de recherche à la fois pour le traitement automatique et des domaines plus linguistiques. Nous avons montré une analyse des erreurs de transcription sur des données journalistiques, une étude similaire pour la parole conversationnelle reste à faire. L'importance du modèle de langage dans bon nombre de confusions d'homophones a été mis en évidence notamment par une mesure modifiée du taux d'erreur de mot. Les ambiguïtés dues aux homophones expliquent une large part des erreurs en parole journalistique. Une partie des erreurs peut cependant être attribuée à une modélisation acoustique inadéquate des mots. En particulier des contractions temporelles mettent en défaut la modélisation des mots comme séquences de phonèmes d'une prononciation standard. Alors que ce problème se pose en termes relativement discrets pour la parole journalistique, il devient flagrant sur la parole conversationnelle. Cette dernière est à débit plus variable, parfois très rapide, parfois très lente, ce qui est illustré par une distribution des durées phonématiques globalement plus étalée, avec une concentration élevée de segments sur la durée minimale de 30 ms. A cet endroit la courbe cumule tous les segments attendus par le modèle de prononciation et qui n'ont pas (ou peu) été réalisés dans la parole effectivement produite. Des résultats d'analyses de corpus présentées ici concernent les fréquences d'occurrences des phonèmes dans les corpus et nous avons pu montrer que ces fréquences varient en fonction du genre de corpus traité et également en fonction du contexte phonémique gauche et droit. Ainsi pour le corpus de données journalistiques nous avons présenté les comptes d'occurrences des voyelles en contextes gauche et droit labial, alvéolaire, postalvéolaire, palato-vélaire et des liquides /l/ et /R/.

Les progrès accomplis en traitement automatique permettent d'aborder bon nombre de recherches sous un angle nouveau. La disponibilité de corpus et d'instruments pour

l'accès au contenu permet de poser un nombre élevé de questions en même temps et d'avoir très vite, si ce n'est une réponse, au moins une tendance. Nous vivons actuellement une révolution technologique qui permettra d'enrichir le domaine de la linguistique de l'oral de nouveaux instruments et de méthodologies expérimentales exploitant de grands corpus [24]. L'ère chomskyenne a rendu pendant des décennies l'usage de corpus en linguistique pour le moins suspect, si ce n'est hors sujet. Sans vouloir rentrer dans des polémiques scientifiques, force est de constater que nous sommes aujourd'hui à un tel point d'accès facile à des données orales qu'il serait non scientifique de refuser l'étude de ces données, dont le corpus ESTER est certainement un exemple important pour le français. De telles études nous pouvons espérer dégager de nouvelles connaissances sur la langue orale et les performances des locuteurs en lien avec la neuro- et psycholinguistique. Ces connaissances seront à terme certainement utiles pour les systèmes de traitement automatique de la parole au sens large, incluant au-delà de la transcription des problématiques comme l'identification des locuteurs, des langues et des accents, la synthèse, la compréhension et le dialogue.

REMERCIEMENTS

Partie des travaux ont été réalisés grâce aux projets interministériel TECHNOLANGUE-ESTER, CNRS TCAN Varcom, ANR PFC-Cor et au projet européen CHIL. Je tiens à remercier ici mes collègues du LIMSI, de la DGA et de Paris 3 qui ont contribué par leurs travaux, réflexions, critiques et suggestions aux résultats présentés.

RÉFÉRENCES

[1] M. Adda-Decker, L. Lamel, "*Pronunciation Variants Across System Configuration, Language and Speaking Style*", Speech Communication "Special Issue on Pronunciation Variation Modeling", **29**, 1999.

[2] M. Adda-Decker, P. Boula de Mareüil, L. Lamel, "*Pronunciation variants in French : schwa & liaison*", 14th International Conference on Phonetic Science, ICPhS-99, août 1999.

[3] M. Adda-Decker, Ph. Boula de Mareüil, G. Adda, & L. Lamel. "Investigating syllabic structures and their variation in spontaneous French", *Speech Communication*, 46 (2005) pp.119-139, Elsevier ed.

[4] M. Adda-Decker, L. Lamel, "Do Speech Recognizers Prefer Female Speakers ?", Eurospeech-Interspeech, Lisbonne, septembre 2005.

[5] C. Barras et al., "Transcriber : development and use of a tool for assisting speech corpora production". Speech Communication, 33(1-2) :5-22, Jan 2001.

[6] C. Blanche-Benveniste, "Constitution et exploitation d'un grand corpus", Rev. Française de linguistique appliquée, 1999, IV-1 (65-74).

[7] L-J. Boë, J-P. Tubach, "Une base de données lexicale orthographique-phonétique du français parlé". Cahiers de grammaire 17, Universite de Toulouse-Le Mirail, novembre 1992.

[8] P. Boula de Mareüil et al., "Liaisons in French : a corpus-based study using morpho-syntactic information". In Proceedings of the International Conference on Phonetic Sciences, ICPhS, Barcelone août 2003.

[9] P. Boula de Mareüil et al. "A quantitative study of disfluencies in French broadcast interviews", dans Proceedings DISS'05, Aix-en-Provence, septembre 2005.

[10] M. Candea et al., "Inter- and intra-language acoustic analysis of autonomous fillers", dans Proceedings DISS'05, Aix-en-Provence, septembre 2005.

[11] P. Delattre (1966) Studies in French and Comparative Phonetics, La Haye, Mouton.

[12] F. Dell (1973). Les règles et les sons : introduction à la phonologie générative. Paris : Hermann. 2e éd.

[13] J.M. Dolmazon et al., *"Organisation de la première campagne AUPELF pour l'évaluation des systèmes de dictée vocale"*, JST97, Avignon, avril 1997.

[14] D. Duez, 2003. Modelling Aspects of Reduction and Assimilation in Spontaneous French Speech, In Proc. IEEE-ISCA Workshop on Spontaneous Speech Processing and Recognition, 2003. Tokyo.

[15] J. Durand, B. Laks, C. Lyche, (2003). Le projet « Phonologie du français contemporain » (PFC). La Tribune Internationale des Langues Vivantes 33 3-9.

[16] P. Encrevé, (1988). La liaison avec et sans enchaînement. Phonologie tridimensionnelle et usages du français. Éditions du Seuil, Paris.

[17] P. Fouché, (1959). Traité de prononciation française. Editions Klincksieck, Paris.

[18] C. Fougeron et al.,"Liaison and schwa deletion in French : an effect of lexical frequency and competition", *Eurospeech*, Aalborg (pp. 639-642), 2001.

[19] S. Galliano et al. "The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News", Eurospeech-Interspeech, Lisbonne, septembre 2005.

[20] J.L. Gauvain et al., "Speaker-Independent Continuous Speech Dictation," Speech Communication, **15** :21-37, Sept. 1994.

[21] J-L. Gauvain et al., "Where Are We In Transcribing French Broadcast News ?", Eurospeech-Interspeech, Lisbonne, septembre 2005.

[22] C. Gendrot, M. Adda-Decker, "Impact of duration on F1/F2 formant values of oral vowels : an automatic analysis of large broadcast news corpora in French and German", Eurospeech-Interspeech, Lisbonne, septembre 2005.

[23] M-A. Hintze, T. Pooley & A. Judge (eds.). "French accents : Phonological and sociolinguistic perspectives", pub. CILT/AFLS, ISBN 1 909031 95 4, 2001.

[24] B. Habert, "Portrait de linguiste(s) à l'instrument", Texto ! Textes et cultures, ISSN 1773-0120, Vol. X, n.4, décembre 2005.

[25] L.F. Lamel et al., "BREF, a Large Vocabulary Spoken Corpus for French," *EuroSpeech'91*.

[26] L.F. Lamel et al. "Issues in Large Vocabulary, Multilingual Speech Recognition," *Eurospeech-95*, Madrid, septembre 1995.

[27] Malécot A. (1974) Frequency of occurence of French phonemes and consonant clusters, Phonetica 29.

[28] N. Nguyen et al., "Detection of liaison consonants in speech processing in French : Experimental data and theoretical implications", Laboratory Approaches to Romance Phonology, édité par P. Prieto et M.J. Solé (John B Benjamins), à paraître.

[29] PRAAT, a system for doing phonetics by computer. Glot International 5(9/10) : 341-345, 2001

[30] I. Vasilescu et al., "Hésitations autonomes dans 8 langues : une étude acoustique et perceptive", Colloque MIDL04 Paris, 29-30 novembre 2004.

[31] C. Woehrling, P. Boula de Mareüil , "Perceptual identification of French varieties in the PFC corpus", 7e Rencontres Internationales du Réseau Français de Phonologie (RFP), Aix-en-Provence, 2005.

[32] S.J. Young et al., "*Multilingual large vocabulary speech recognition : the European SQALE project*", Computer Speech & Language, vol. 11, nb.1, janv. 1997.

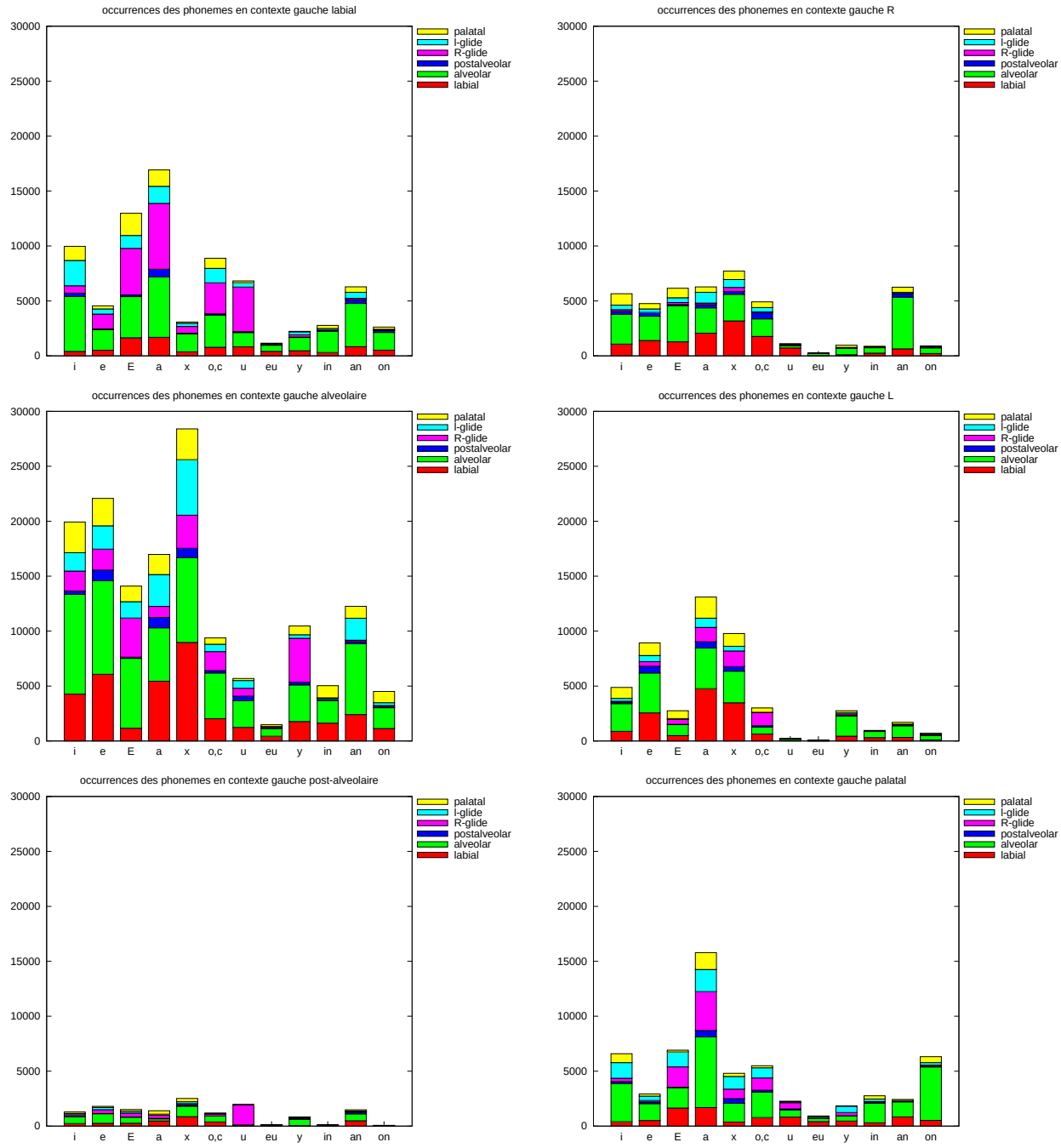

FIG. 10: Fréquences d'occurrence des voyelles en contextes gauche et droit de labiales, d'alvéolaires, de postalvéolaires, de la liquide /R/, de la liquide /l/ et de palato-vélaires mesurées sur le corpus **journalistique**. **En haut à gauche :** le contexte gauche correspond aux **labiales** ; **à droite :** le contexte gauche correspond à **/R/** ; **milieu à gauche :** le contexte gauche correspond aux **alvéolaires** ; **à droite :** le contexte gauche correspond à **/l/** ; **en bas à gauche :** le contexte gauche correspond aux **postalvéolaires** ; **à droite :** le contexte gauche correspond aux **palato-vélaires**. Pour chaque contexte gauche les histogrammes indiquent la distribution des voyelles en fonction des 6 classes de consonnes en contexte droit.