Systèmes d'alignement automatique & études de variantes de prononciation

Martine Adda-Decker et Lori Lamel

Groupe Traitement du Langage Parlé
LIMSI-CNRS, BP 133, 91403 Orsay cédex, FRANCE
{lamel,madda}@limsi.fr
http://www.limsi.fr/TLP

ABSTRACT

This contribution aims at evaluating the use of pronunciation variants across different system configurations and speaking styles in French. The study is limited to the use of variants during speech alignment, given an orthographic transcription and a phonemically represented lexicon, thus focusing on the modeling abilities of the acoustic word models. Parallel and sequential variants are tested in order to measure the spectral and temporal modeling accuracy. To measure the need for variants we have defined the *variant2+* rate which is the percentage of words in the corpus, not aligned with the most common phonemic transcription. Alignment results using different acoustic model sets demonstrate the dependency between acoustic model accuracy and pronunciation variants. A comparison between read and spontaneous speech is presented for French based on alignments from BREF (read) and MASK (spontaneous) data.

1. INTRODUCTION

Les variantes de prononciation peuvent s'expliquer pardifférents facteurs, comme le style de parole, la vitesse d'élocution, des habitudes individuelles ou des accents régionaux... La modélisation de variantes de prononciation pour la reconnaissance automatique de la parole a attiré beaucoup d'intérêt ces dernières années [Spe99] et des exemples pour le français peuvent être trouvés dans [Per98, Mok98], une étude sur l'influence de la vitesse d'élocution dans [Lus98]. L'ajout de variantes de prononciation dans le dictionnaire du système de reconnaissance permet d'accroître les possibilités de modélisation acoustique des mots, et l'effet souhaité est d'arriver à des modèles de mot plus précis et par là à un meilleur décodage. Cependant si les variantes rajoutées ne sont pas pertinentes pour les données acoustiques traitées et/ou par rapport aux faiblesses du décodeur, les performances globales du système peuvent décroître. Combien de fois a-t-on pu observer que les variantes ajoutées n'ont pas permis d'arriver à une amélioration globale: alors que des erreurs sont ponctuellement corrigées, de nouvelles erreurs peuvent s'introduire. Les variantes contribuant à augmenter le taux d'homophones dans le système, elles deviennent des sources d'erreurs potentielles et elles ne sont que peu utilisées dans nos systèmes de reconnaissance [Lam96].

Dans une étude récente [Add99] nous nous sommes intéressés aux variantes possibles lors de simples expériences d'alignement où les dictionnaires de prononciations contiennent un nombre plus ou moins élevé de variantes et où l'alignement n'est guidé que par les modèles acoustiques sans biais du modèle de langage. Nous continuons ici ce travail avec différents corpus en français. L'utilité des variantes est mesurée suivant différents axes: la configuration du système et le style de parole dans les corpus (lu, spontané). On distingue des variantes de prononciation séquentielles et parallèles. Les variantes séquentielles permettent certains phonèmes d'être optionnels ce qui donne une plus grande flexibilité pour la modélisation temporelle des mots. Les variantes parallèles permettent de remplacer un phonème par n'importe quel autre phonème d'un sous-ensemble défini a priori, augmentant ainsi les possibilités de modélisation spectrale. Les variantes observées lors de l'alignement peuvent s'expliquer soit simplement par des faiblesses de modélisation oubien, si la modélisation acoustique est précise, les variantes correspondent à une réalité linguistique et peuvent servir à des études phonétiques.

2. CORPUS DE PAROLE

Deux corpus sont utilisés pour nos expériences. La parole lue provient du corpus BREF [Lam91] qui correspond à la lecture d'articles du journal *LeMonde*. La parole spontanée concerne des demandes d'informations SNCF et a été enregistrée au LIMSI pour le projet ESPRIT MASK (Multimodal-Multimedia Automated Service Kiosk) [Lam95]. Le contenu de ces corpus est résumé dans la Table 1.

Table 1: Pour chaque corpus sont indiqués le nombre d'énoncés, la durée de parole en nombre d'heures, le nombre total de mots et le nombre de mots distincts.

Corpus	BREF	MASK
style	lu	spontané
#énoncés (distincts)	6.5k	38k
durée parole	120h	35h
#mots(total)	1.1M	260k
#mots(distincts)	25k	2k

La figure 1 montre, pour chaque corpus, la couverture lexicale cumulée en fonction du rang de fréquence des mots. Pour MASK (parole spontané, limitée à un domaine) les 10 mots les plus fréquents représentent 30% des mots du corpus, alors que pour les journaux lus ils couvrent 20%. Avec les 100 mots les plus fréquents 80% des mots du corpus MASK sont couverts, mais seulement 50% pour BREF. Alors que la courbe de BREF est quasi-linéaire sur une échelle logarithmique, la courbe de MASK a une forte pente entre les rangs 10 et 200 et s'aplanit rapidement audelà, ce qui traduit la grande spécificité du corpus.

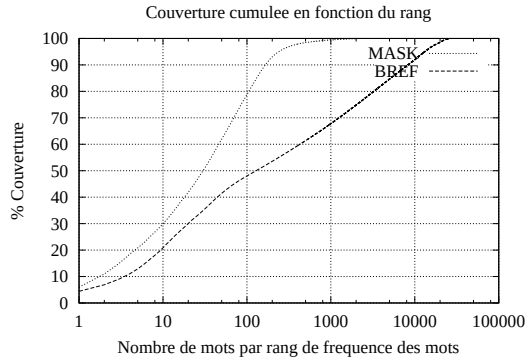

Figure 1: Couverture lexicale en fonction du nombre de mots triés par rang de fréquence du mot pour les corpus spontané (MASK) et lu (BREF).

3. DICTIONNAIRES DE PRONONCIATION

À partir de nos dictionnaires de prononciation standards (référence) nous avons créé des dictionnaires augmentés de variantes parallèles ou séquentielles. Notre but est d'accroître notre compréhension des facultés et limites des modèles acoustiques concernant la modélisation spectrale et temporelle en comparant des résultats d'alignement à travers différents styles de parole.

Dictionnaires de référence Dans la Table 2 on peut voir quelques entrées de notre dictionnaire de prononciation de référence utilisé pour l pprentissage des modèles acoustiques. Ces dictionnaires contiennent typiquement de 10 à 20% de variantes concernant les mots outils fréquents, les nombres ((4) dans la Table 2), les acronymes (5) et les noms propres (6) souvent d'origine étrangère. Un nombre important de variantes concernent le schwa en position finale (3,4) et les liaisons (2,4).

Table 2: Exemples d'entrées lexicales dans le dictionnaire de référence avec variantes parallèles ([ ]: phonèmes au choix) et séquentielles ({ }: phonèmes optionels).

république	repyblik	(1)
les	le{z}	(2)
prendre	prɑ̃dr{ə} prɑ̃d	(3)
dix	dis{ə} di{z}	(4)
DM	d[œ,ə]tSmark dɛm	(5)
Morgan	mɔrgɑ̃ mɔrgɑn	(6)

Dictionnaires à variantes séquentielles Des dictionnaires incluant un très grand nombre de variantes séquentielles ont été dérivées à partir des dictionnaires de référence en rendant une partie des phonèmes optionnels. Ces dictionnaires, appelés *Vopt* (voyelles optionnelles) et *Copt* (consonnes optionnelles) ont pour but de localiser d'éventuels problèmes de modélisation temporelle concernant les modèles acoustiques de mots. Un extrait de ces dictionnaires est montré dans la Table 3. Le phénomène du schwa optionnel en fin de mot est ainsi pris en compte dans les dictionnaire séquentiels. Le dictionnaire *Copt* peut servir à étudier les phénomènes de réduction concernant les clusters de consonnes. Le dictionnaire *Vopt* peut servir à étudier si d'autres voyelles que le schwa sont susceptibles de disparaître et dans quel contexte. De telles omissions, a priori rares, apparaissent assez fréquemment en spontané, accompagnées d'une restructuration syllabique.

Table 3: Exemples d'entrées lexicales dans les dictionnaire *Vopt* et *Copt* montrant une grande flexibilité temporelle. Le schwa final {ə} est optionnel dans tous les dictionnaires.

Vopt	république	r{e}p{y}bl{i}k{ə}
Copt	république	{r}e{p}y{b}{l}i{k}{ə}

Dictionnaires à variantes parallèles Ces dictionnaires ont été générés en définissant des classes de phonèmes et en autorisant un phonème d'une classe donnée à être remplacé par n'importe quel autre membre de cette même classe. Pour chaque classe un dictionnaire spécifique a été créé. La Table 4 montre les classes de phonèmes utilisés dans les travaux décrits ici.

Table 4: Classes de phonèmes pour dictionnaires parallèles.

Vclass1	ɛ e	*Cclass1*	b d g v
Vclass2	ɛ̃ ə œ ɔ	*Cclass2*	l r ɥ w j

En français beaucoup de quasi-homophones se différencient par la caractéristique ouvert/fermé de voyelles (e.g.: est /ɛ/, et /e/). Dans la parole courante cette distinction peut disparaître, l'identification correcte s'appuyant davantage sur le contexte que sur le signal acoustique.

Pour donner une indication de la complexité des différents dictionnaires nous indiquons dans la table 5 le rapport du nombre total de variantes dans le dictionnaire par le nombre total d'entrées lexicales. Les dictionnaires *Copt* admettent le plus de variantes. Les taux globalement plus élevées pour BREF s'expliquent simplement par un nombre plus élevé d'entrées lexicales plus longues. Pour les variantes parallèles peu de phonèmes peuvent être modifiés et les taux, relativement faibles, sont les plus forts pour la classe *Cclass2* des liquides et glissantes.

Table 5: Rapports $\frac{\#variantes}{\#entres}$ dans les dictionnaires références, séquentiels *Vopt*, *Copt*, parallèles *Vclass*, *Cclass*.

	MASK	*BREF*
Référence	1.1	1.2
Vopt	9.5	17.3
Copt	20.0	33.7
Vclass1	1.7	2.5
Vclass2	2.4	4.0
Cclass1	2.7	4.3
Cclass2	10.1	15.1

4. MESURE: LE TAUX DE *Variant2+*

Pour mesurer l'utilité des variantes lors de l'alignement automatique nous comptons la proportion de mots alignés avec les prononciations minoritaires. Cette mesure, appelée le taux de *Variant2+* [Add99], donne le pourcentage de mots alignés avec des variantes de rang de fréquence $r_\varphi \geq 2$. Ce taux donne une indication sur le besoin de variantes pour une meilleure modélisation acoustique oubien, de manière équivalente, sur la capacité d'une prononciation unique de rendre compte de toutes les occurrences de

ce mot. Le taux de *Variant2+* est définie dans les équations [1,2], où n désigne le mot de rang de fréquence n, $\#occ_n$ le nombre d'occurrences du mot n dans le corpus et $\#align_n^{r_\varphi=1}$ le nombre de mots alignés avec la variante majoritaire.

$$\%var2+_n = 100 \times (\#occ_n - \#align_n^{r_\varphi=1})/\#occ_n \quad [1]$$

$$\%Variant2+(n) = \frac{\sum_{i=1}^{n} var2+_i}{n} \quad [2]$$

La mesure $var2+_n$ ([1]) est spécifique au mot de rang n et le taux $Variant2+(n)$ ([2]) intègre tous les mots du rang 1 au rang n. Le taux de *Variant2+* **global** est le *%Variant2+*(**N**), avec **N** la taille du lexique.

Table 6: Exemple d'entrée lexicale n dans le dictionnaire *référence*, nombre d'occurrences dans BREF ($\#occ_n$), ltaux $var2+$ et détail des différentes variantes triées par rang de fréquence r_φ avec le nombre d'occurrences alignées ($\#align_n$).

entry n	$\#occ_n$	$var2+_n$	*phon.*	r_φ	$\#align_n$
les	21362	24%	le	1	16262
			lez	2	5100

Les figures 2-8 donnent le taux de *Variant2+* en fonction du rang de fréquence des mots. Dans chaque figure la courbe obtenue avec le dictionnaire de référence (du système de reconnaissance) est rajoutée, ce qui permet d'évaluer la proportion de mots mieux modélisés par les dictionnaires de variantes qu'avec les dictionnaires de reconnaissance.

5. CONDITIONS & RÉSULTATS

5.1. Configurations d'alignement

Des expériences d'alignement automatique ont été faites avec différents modèles acoustiques indépendants du contexte (CI: 36, 35) et dépendants du contexte (CD: 637, 594) pour MASK et BREF respectivement. Pour BREF un deuxième ensemble de 761 modèles CD a été utilisé.

La table 7 montre que le taux de *Variant2+* obtenus avec les dictionnaires *Vopt* et *Copt* décroît de manière significative en passant de modèles CI à des modèles CD. Cette même observation a pu être faite avec tous les dictionnaires de prononciation. Ce résultat, vérifié aussi sur l'anglais [Add99], montre que le besoin de variantes diminue si le nombre de modèles acoustiques augmente.

Table 7: Taux de *Variant2+* global pour différents modèles acoustiques et dictionnaires.

dictionnaires	mod. ac.	*MASK*	*BREF*
Vopt	*CI*	22.2	18.6
	CD	13.0	14.8
Copt	*CI*	27.0	21.0
	CD	14.5	16.2

5.2. Styles de parole et types de variantes

Dans les figures 2 et 3 le taux de *Variant2+* obtenu avec les dictionnaires *Vopt* sur la parole lue et spontanée est donné en fonction de la fréquence des mots. La figure 4 regroupe les mêmes informations pour les dictionnaires *Copt*.

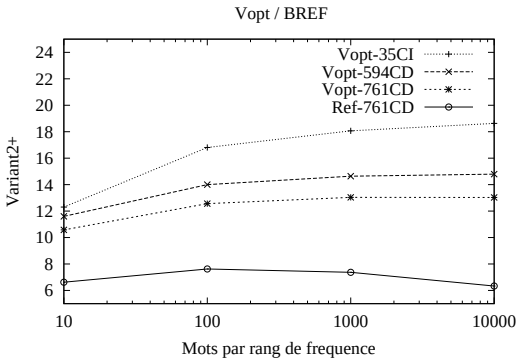

Figure 2: Taux de *Variant2+* en fonction du rang des mots pour la parole lue (BREF) avec le dictionnaire *Vopt* et différents modèles acoustiques.

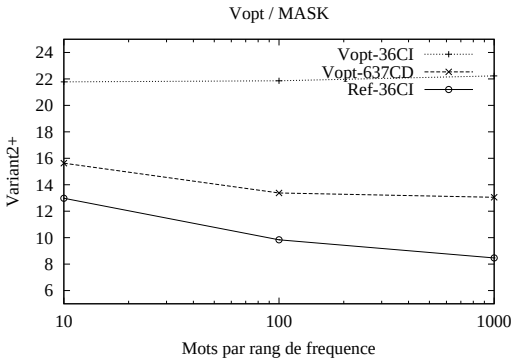

Figure 3: Taux de *Variant2+* en fonction du rang des mots pour la parole spontanée (MASK) avec le dictionnaire *Vopt* et différents modèles acoustiques.

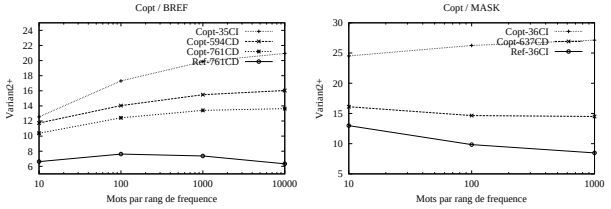

Figure 4: Taux de *Variant2+* en fonction du rang des mots pour BREF et MASK utilisant les dictionnaires *Copt* et différents modèles acoustiques.

Des courbes similaires sont données pour les dictionnaires *Vclass* and Cclass dans les figures 5- 8. Le taux de *Variant2+* est très élevé pour les mots fréquents en parole spontanée, la même chose n'est pas vrai pour la parole lue. Les dictionnaires *Vopt* et *Copt* admettent plus de variantes en parole spontanée qu'en parole lue, spécialement avec des modèles CI, mais les modèles CD permettent de réduire considérablement ce taux. Les modèles CD, appris avec les dictionnaires de référence, absorbent une part importante de ces variantes séquentielles, surtout pour la parole spontanée. On peut donc faire l'hypothèse que beaucoup de modèles de phone en contexte représentent des segments acoustiques différents d'un simple segment phonétique. L'analyse des résultats obtenus avec les dictionnaires de classes de consonnes (figures 5, 6) montrent que les modèles acoustiques des consonnes sont relativement précis. Pour la parole lue les taux de *Variant2+*

restent faibles même pour les modèles CI, ce qui n'est pas le cas pour la parole spontanée. Les modèles CD ramènent les courbes très près des courbes référence. Pour les classes de voyelles (figures 7 et 8) des variations plus importantes peuvent être mesurées. La classe *Vclass1* a un taux important de *Variant2+* avec une forte proportion de substitutions ɛ→e. La comparaison entre parole lue et spontanée (BREF et MASK) montre qu'avec les modèles acoustiques CI, la parole spontanée admet significativement plus de variantes que la parole lue et que l'emploi de modèles CD réduit toujours le taux de *Variant2+*: ceci est particulièrement vrai pour MASK où le vocabulaire limité fait que les modèles de phones contexte-dépendant deviennent vite mot-dépendent.

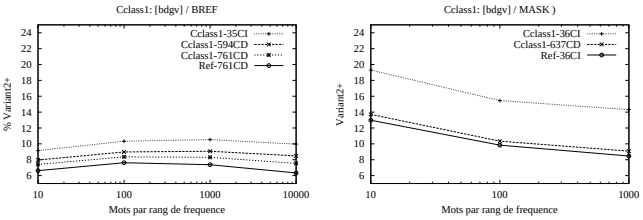

Figure 5: Taux de *Variant2+* en fonction du rang pour BREF and MASK avec les dictionnaires de *Cclass1* ([bdgv]).

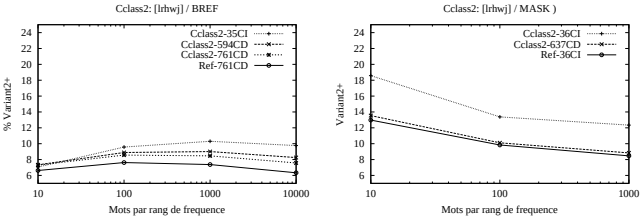

Figure 6: Taux de *Variant2+* en fonction du rang pour BREF et MASK avec les dictionnaires de *Cclass2* ([lrɥwj]).

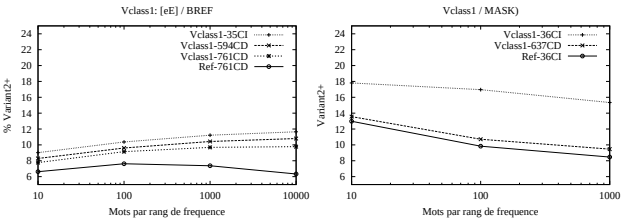

Figure 7: Taux de *Variant2+* en fonction du rang pour BREF et MASK avec les dictionnaires de *Vclass1* ([eE]).

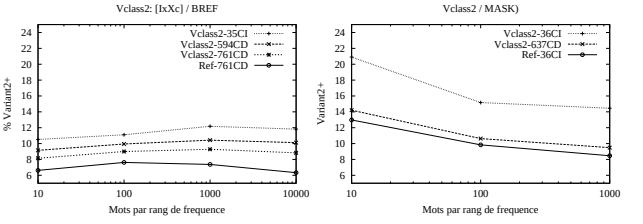

Figure 8: Taux de *Variant2+* en fonction du rang pour BREF et MASK avec les dictionnaires de *Vclass2* ([ɛ̃ ə œ ɔ]).

À partir de ces mesures globales beaucoup d'investigations précises sont possibles. Par exemple nous avons examiné les clusters plosives-liquides en position finale de mot où un taux relativement fort de variation séquentielle peut être supposé. Dans le corpus BREF environ 25k mots sont concernés, 7k dans MASK. Des taux de *Variant2+* de 38% et de 51% sont obtenus avec des modèles CI et les dictionnaires *Copt* pour BREF et MASK respectivement.

6. DISCUSSION & PERSPECTIVES

Des résultats comparatifs d'alignement utilisant différents ensembles de modèles acoustiques sur différents types de corpus avec des dictionnaires de pronnociations à taux de variantes élevés montrent que le besoin de variantes de prononciation dépend fortement de la précision des modèles acoustiques. Augmenter le nombre de modèles contexte-dépendants, couvrant progressivement plus de contextes diminue le besoin de variantes.

Les dictionnaires de variantes séquentielles montrent que si le choix est donné au système d'alignement, un pourcentage important des mots sont alignés avec un nombre de phonèmes différents du nombre prévu. En français la liaison et le schwa final optionnel sont des phénomènes bien connus permettant de générer un nombre variable de phonèmes par mot [Add99b]. Les modèles dépendants du contexte permettent d'absorber une partie de cette variabilité, particulièrement si ces modèles sont appris et utilisés sur un même vocabulaire de taille réduite (MASK, spontané). Les modèles deviennent ainsi plus spécifiques au mot et moins spécifiques au phonème.

Cette étude nous permet d'analyser le comportement global des modèles acoustiques de phones lors de l'alignement avec des dictionnaires très permissifs permettant de simuler localement la reconnaissance phonétique à un prix de calcul beaucoup plus faible. Cette approche permet de focaliser l'attention sur des problèmes précis qui peuvent se poser autant d'un point de vue d'ingénierie de la parole que d'un point de vue linguistique.

BIBLIOGRAPHIE

[Spe99] **Speech Communication** "Special Issue on Pronunciation Variation Modeling", **29**, 1999.

[Lam96] L.Lamel, G.Adda, "On Designing Pronunciation Lexicons for LVCSR", *ICSLP'96*.

[Per98] G. Pérennou, L. Brieussel-Pousse, "Phonological Component in Automatic Speech Recognition", **ESCA**-ETRW Pronunciation Modeling for ASR, Rolduc, May 1998.

[Lus98] E. Fossler-Lussier, N. Morgan, "Effects of speaking rate and word frequency on conversational pronunciations", **ESCA**-ETRW Pronunciation Modeling for ASR, Rolduc, May 1998.

[Add99] M. Adda-Decker, L. Lamel, *"Pronunciation Variants Across Systems, Languages and Speaking Style,"*, Speech Communication, **29**, pp.83-99, 1999.

[Mok98] H. Mokbel, D. Jouvet, "Derivation of the optimal phonetic transcription set for a word from its acoustic realisations", **ESCA**-ETRW Pronunciation Modeling for ASR, May 1998.

[Add99b] M. Adda-Decker, P. Boula de Mareuil, L. Lamel, "Pronunciation variants in French: schwa & liaison", ICPhS-99, août 1999.

[Lam95] L. Lamel et al., "Development of Spoken Language Corpora for Travel Information", *EuroSpeech'95*.

[Lam91] L.F. Lamel, J.L. Gauvain, M. Eskénazi, "BREF, a Large Vocabulary Spoken Corpus for French," *EuroSpeech'91*.