METHODES ET OUTILS POUR L'ANALYSE ACOUSTIQUE DES SYSTEMES VOCALIQUES

Version 1.0 (janvier 2004)

Noël Nguyen et Robert Espesser[1]

0. Introduction

L'objectif général de ce travail est de procéder à une analyse acoustique détaillée des voyelles du français à partir des enregistrements réalisés dans le cadre du projet PFC. Il vise à caractériser la structure acoustique des voyelles et à établir des comparaisons dans ce domaine entre les systèmes vocaliques relatifs à différents points d'enquête. Dans une première étape, nous nous sommes focalisés sur les voyelles dans les listes de mots lus isolément.

Dans cet article, nous présentons le matériel utilisé, les paramètres acoustiques mesurés et les procédures d'analyse que nous avons employées. L'accent est placé sur deux points: 1) les procédures d'extraction semi-automatique des données nécessaires à l'exploitation à grande échelle de la base PFC; 2) les procédures de normalisation utilisables pour minimiser les variations liées aux différences anatomiques entre locuteurs.

1. Matériel

Le matériel se compose de 46 mots tirés de la liste PFC (voir tableau 1). Ces mots sont rassemblés dans le protocole en cinq grandes familles (A, é/è, EU, O, et nasales), et ils sont destinés à permettre de caractériser le système vocalique du locuteur au niveau phonémique. Pour une majorité d'entre eux, ces mots sont prononcés une fois par chaque locuteur (10 le sont à deux reprises). Dans les mots plurisyllabiques, la position de la syllabe contenant la voyelle-cible varie selon les mots.

mot	rép.	position voy. cible	étiquette
mal	1	V1	A
mâle	1	V1	A
malle	1	V1	A
pâte	2	V1	A
patte	2	V1	A
ras	1	V1	A
rat	1	V1	A
épais	2	V2	é/è
épée	2	V2	é/è
épier	1	V2	é/è
étrier	1	V3	é/è
étriller	1	V3	é/è
liège	1	V1	é/è
lierre	1	V1	é/è
pécheur	1	V1	é/è
pêcheur	1	V1	é/è
piquais	1	V2	é/è
piqué	1	V2	é/è
piquer	1	V2	é/è
piquet	1	V2	é/è

mot	rép.		étiquette
jeune	2	V1	EU
jeûne	2	V1	EU
creuse	1	V1	EU
creux	1	V1	EU
dégeler	1	V2	EU
déjeuner	1	V2	EU
des genêts	1	V2	EU
des jeunets	1	V2	EU
feutre	1	V1	EU
meurtre	1	V1	EU
peuple	1	V1	EU
beauté	2	V1	O
botté	2	V1	O
paume	1	V1	O
pomme	1	V1	O
rauque	1	V1	O
Roc	1	V1	O
rhinocéros	1	V4	O
blanc	1	V1	N
blond	1	V1	N

[1] Laboratoire Parole & Langage, UMR 6057, CNRS & Université de Provence 29 av Robert Schuman, 13621 Aix-en-Provence.
e-mail: nguyen@lpl.univ-aix.fr, espesser@lpl.univ-aix.fr

fête	1	V1	é/è
fêtard	1	V1	é/è
fêter	1	V1	é/è
faites	1	V1	é/è

brin	2	V1	N
brun	2	V1	N

Tableau 1 : Liste des mots utilisés, nb de répétitions, position de la voyelle-cible à l'intérieur du mot porteur, étiquette associée à la voyelle

2. Procédures d'extraction des données

Les mesures sont réalisées selon une procédure semi-automatique offrant un bon compromis entre la vitesse de traitement et la précision des valeurs obtenues. Cette procédure se présente sous la forme d'un ensemble de modules de traitement implémentés sur une station de travail linux. Les données initiales sont constituées par les fichiers acoustiques au format .wav. À chaque fichier est associée une transcription orthographique de la séquence prononcée. Cette transcription orthographique est convertie en une transcription phonémique au moyen d'un dictionnaire. Un aligneur automatique est ensuite utilisé pour découper le signal de parole en une suite de segments associés chacun à un phonème. L'alignement phonèmes-signal est vérifié au moyen du système MES, un éditeur de signal développé par le deuxième auteur (http://www.lpl.univ-aix.fr/ext/projects/mes_signaix.htm/). Une extraction automatique des formants est ensuite réalisée à partir du signal acoustique au moyen du système ESPS (Entropic). Les fréquences de F1 et de F2 sont relevées au milieu de la voyelle-cible et stockées dans un fichier. Les valeurs sont en partie manuellement vérifiées dans MES.

2.1 L'alignement automatique phonèmes-signal

L'aligneur employé a été développé au LORIA par Fohr et Laprie (http://www.loria.fr/equipes/parole/). Il est basé sur un ensemble de modèles de Markov cachés (8 gaussiennes). Une liste des phonèmes reconnus par le système est présentée dans le tableau 2, avec les symboles utilisés correspondants. La figure 1 illustre la façon dont la séquence « des jeunets » (locuteur GM, Douzens) a été automatiquement segmentée au moyen de l'aligneur.

aligneur	p	t	k	b	d	g	f	s	S	v	z	Z	m	n	w	j	l	R
API	p	t	k	b	d	g	f	s	ʃ	v	z	ʒ	m	n	w	j	l	χ, ʁ

aligneur	i	y	e	2	@	9	A	o	u	U~	a~	o~
API	i	y, ɥ	e, ɛ	◌̩	◌́	◌̥	a, ɑ	o, ø	u	ɛ̃, ◌̃	ɑ̃	ɔ̃

Tableau 2: Liste des consonnes et des voyelles reconnues par l'aligneur

Pour évaluer la précision avec laquelle l'aligneur se montre capable de localiser l'emplacement des frontières entre phonèmes, nous avons procédé à des comparaisons entre cet étiquetage automatique et un étiquetage manuel. La base de données acoustiques utilisée a été mise en place pour les besoins d'une étude sur l'harmonie vocalique en français (Nguyen & Fagyal, 2003) et elle est semblable à la base PFC pour les points qui nous concernent ici. Cette base se compose de 276 mots disyllabiques enregistrés chacun à quatre reprises par quatre locuteurs (3 femmes, 1 homme) représentant deux accents régionaux. Pour chaque voyelle à l'intérieur de chaque mot, nous avons calculé l'intervalle entre l'étiquette automatique et l'étiquette manuelle en début de voyelle d'une part, et en fin de voyelle d'autre part.

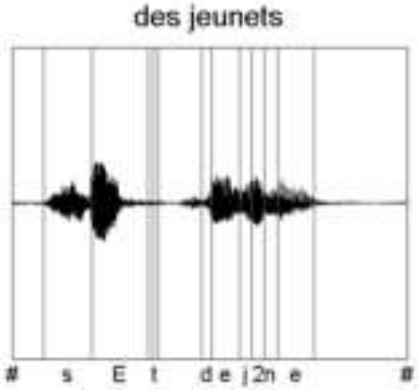

Figure 1: Segmentation automatique de la séquence « des jeunets »

La figure 2 illustre la taille des écarts automatique/manuel. Elle représente la manière dont les étiquettes manuelles sont distribuées vis-à-vis des étiquettes automatiques au début et à la fin de V1 et de V2 (de la gauche vers la droite). L'intervalle est négatif lorsque l'étiquette manuelle a été positionnée à gauche de l'étiquette automatique dans le signal, et il est positif dans le cas contraire. La largeur de chaque barre sur les histogrammes est équivalente à une durée de 10 ms. La figure montre que la congruence entre étiquettes automatique et manuelle est bonne en début de voyelle et moins bonne en fin de voyelle, notamment en ce qui concerne V2. On constate par ailleurs que l'étiquette manuelle a tendance à être située à gauche de l'étiquette automatique en début de voyelle, et à droite en fin de voyelle. (Cela signifie que la durée estimée de la voyelle serait en moyenne plus courte si elle devait être calculée à partir des étiquettes automatiques, par opposition aux étiquettes manuelles.)

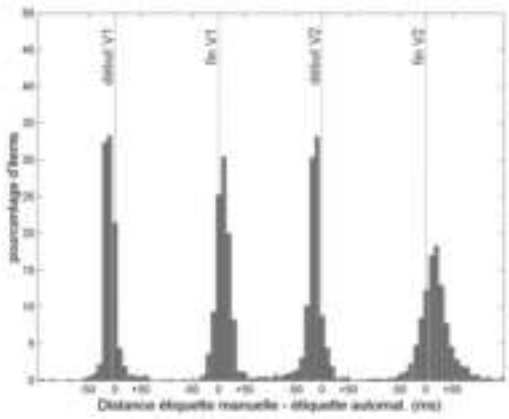

Figure 2 : Distribution des étiquettes manuelles de début et de fin de voyelle vis-à-vis des étiquettes automatiques correspondantes

Il est possible à partir des étiquettes initiales de construire une nouvelle étiquette coïncidant avec le milieu acoustique de chaque voyelle. La figure 3 représente la distribution des étiquettes manuelles vis-à-vis des étiquettes automatiques en ce point-là pour V1 et V2. Les écarts automatique/manuel observés en début et en fin de voyelle se compensent, et chaque distribution est centrée sur la position de l'étiquette automatique (ce qui signifie que dans une majorité des cas, l'intervalle entre étiquettes manuelle et automatique est nul). Comme la figure 2 le laissait présager, les écarts sont plus grands pour V2 que pour V1.

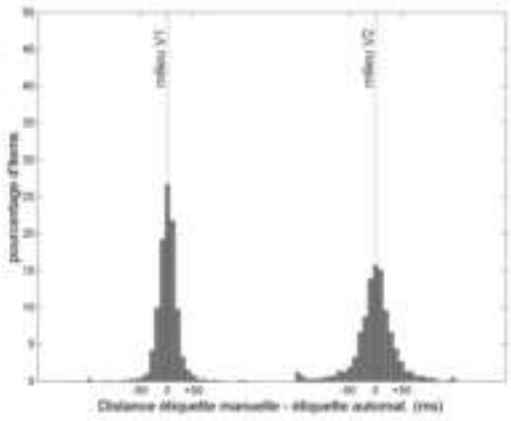

Figure 3 : Distribution des étiquettes manuelles en milieu de voyelle vis-à-vis des étiquettes automatiques correspondantes

Les écarts entre étiquettes automatiques et manuelles sont selon toute vraisemblance conditionnés par le contexte phonétique dans lequel la voyelle se présente. Nous avons cherché à mieux caractériser l'influence de ce contexte en fin de voyelle (puisque les écarts sont plus importants en ce point qu'en début de voyelle). Les résultats sont présentés dans le tableau 3. Les différents éléments pouvant faire suite à V1 et à V2 sont ici rangés en fonction de la valeur de l'écart (qui augmente du haut vers le bas). La valeur indiquée ici désigne l'écart maximum pour 75% des valeurs. Pour prendre un exemple, l'écart automatique/manuel est inférieur ou égal à 8 ms dans 75% des cas à la fin de V1 lorsque celle-ci est suivie par un /l/. Ce tableau nous permet d'identifier les contextes dans lesquels les écarts sont les plus importants (/j/ pour V1 par ex.).

V1	
contexte droit	quartile 75%
/l/	8 ms
/m/	9 ms
occl. +voisée	13 ms
/n/	13 ms
/v/	14 ms
occl. –voisée	18 ms
/f/	20 ms
/s/	22 ms
/r/	22 ms
/ʃ/	24 ms
/z/	25 ms
/j/	34 ms

V2	
contexte droit	quartile 75%
/n/	17 ms
Fin de fichier ou occl. –voisée	21 ms
/l/	21 ms
/s/	21 ms
/f/	21 ms
/ʃ/	25 ms
/ʒ/	33 ms
# (pause)	38 ms
/r/	46 ms
/z/	51 ms
/j/	61 ms

Tableau 3 : Amplitude des écarts automatique/manuel en fin de voyelle, selon le contexte droit, pour V1 et V2

En résumé, il ressort de ces premières évaluations que le milieu de la voyelle est localisé par l'aligneur avec une bonne précision générale (telle que celle-ci peut être estimée par comparaison avec un étiquetage manuel) dans la mesure où l'écart automatique/manuel est inférieur à 20 ms dans 75% des cas. Cependant, il est clair que la précision de l'alignement varie en fonction de la position de la voyelle dans le mot et du contexte droit, en particulier.

Il est difficile de déterminer a priori quel sera l'impact de ces écarts dans la localisation du milieu de la voyelle, sur les fréquences estimées de F1 et de F2. Cela dépend de la vitesse à laquelle F1 et F2 varient en fréquence au voisinage de ce point. Si les formants présentent une trajectoire plate dans cette partie de la voyelle, les fréquences relevées seront insensibles à de petits écarts dans l'emplacement estimé du milieu de voyelle. À l'inverse, si les formants présentent des variations rapides, les fréquences relevées seront évidemment différentes selon la position attribuée au milieu de voyelle.

Des analyses plus approfondies sont en cours de réalisation. Dans l'intervalle, nous préconisons une solution mixte, dans laquelle l'étiquetage automatique réalisé par l'aligneur est soumis à une vérification manuelle pour les voyelles courtes, et les voyelles en position finale et/ou devant /z/, /r/ et semi-voyelle.

2.2 Mesures de formant

Un suivi automatique des formants sur toute la durée du signal acoustique est réalisé au moyen de la fonction formant (ESPS/Entropic). Le signal est soumis à un filtrage passe-bas (fréqu. de coupure: 10 kHz) à un filtrage passe-haut (fréqu. de coupure: 80 Hz), et il est préemphasé (coeff. de préemphase : 0.94). La fenêtre temporelle d'analyse est du type $\cos^4$, d'une durée de 49 ms et elle est déplacée dans le signal par pas de 5 ms. L'analyse spectrale est du type LPC (autocorrélation, 12 coefficients de prédiction). Les valeurs de fréquence pour F1 et F2 au milieu de la voyelle-cible sont ensuite isolées et rangées dans un fichier.

Les erreurs de détection potentielles sont identifiées à partir d'un examen visuel portant sur la manière dont les voyelles se distribuent dans le plan F1-F2. Des vérifications sont réalisées dans MES, à partir du spectre LPC et du spectre FFT correspondant, et d'un spectrogramme à bande large. Le tableau 4 indique le pourcentage de valeurs de fréquence manuellement vérifiées par locuteur pour deux points d'enquête, Douzens et Tournai.

Douzens	
locuteur	%
al	11
gm	18
jp	4
ld	18
tg	23
dp	20
mg1	20
mg2	11
ml	9
nb	9
moyenne	**14**

Tournai	
locuteur	%
bd	5
cb	20
cw	18
fb	5
jl	13
mp1	14
mp2	11
nh	5
ol	14
pc	5
pm	7
tm	16
moyenne	**11**

Tableau 4 : % de valeurs de fréquence pour F1 et F2 manuellement vérifiées.

3. Points d'enquête analysés

Les analyses accomplies jusqu'à présent ont porté sur deux points d'enquête, Douzens et Tournai. Les enregistrements ont été réalisés à Douzens par Jacques Durand et collaborateurs, et à Tournai par Philippe Hambye et collaborateurs. D'autres analyses sont en cours de réalisation sur les enregistrements effectués en Vendée par Géraldine Mallet et collaborateurs. Les locuteurs ont été divisés en trois classes d'âge (classe 1 : moins de 30

ans ; classe 2 : entre 30 et 50 ans ; classe 3 : plus de 50 ans). La répartition des locuteurs en fonction de l'âge et du sexe est présentée dans le tableau 5 pour les deux points d'enquête.

Classe d'âge	Douzens		Tournai	
	H	F	H	F
1	1	2	3	2
2	1	2	2	2
3	3	1	1	2

Tableau 5 : Effectifs en fonction de l'âge et du sexe pour les deux points d'enquête.

4. Procédures de normalisation inter-individuelle

La figure 4 représente la distribution des voyelles en fonction de la fréquence de F1 (axe vertical) et de F2 (axe horizontal), pour les deux points d'enquête, Douzens et Tournai. Les valeurs augmentent vers le bas pour F1 et vers la gauche pour F2, de telle sorte que les voyelles antérieures fermées se trouvent en haut à gauche, comme dans une carte vocalique traditionnelle. Les ellipses sont associées chacune à une catégorie vocalique : e, ɛ, ø, œ, a, ɔ, o, de la gauche vers la droite. La taille de l'ellipse est de deux écarts types sur chacun de ses deux axes. L'orientation de ces axes a été établie au moyen d'une analyse en composantes principales. On constate que les ellipses sont orientées vers la gauche et (dans une moindre mesure) vers le bas. Ces variations pour chaque voyelle sont en partie attribuables aux différences anatomiques individuelles (puisque la fréquence moyenne de F1 et de F2 est plus élevée lorsque le conduit vocal du locuteur est plus court).

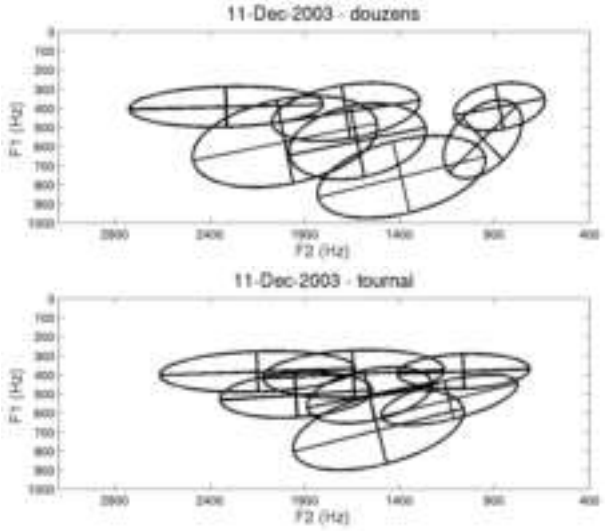

Figure 4 : Distribution des voyelles dans le plan F1-F2 pour les deux points d'enquête

La figure 5 représente la fréquence moyenne de F1 et F2 calculée sur l'ensemble des voyelles, en fonction de la classe d'âge et du sexe, pour les deux points d'enquête. Comme cela était prévisible, on constate d'importantes différences entre hommes et femmes, avec un écart de 150 à 200 Hz en faveur des femmes.

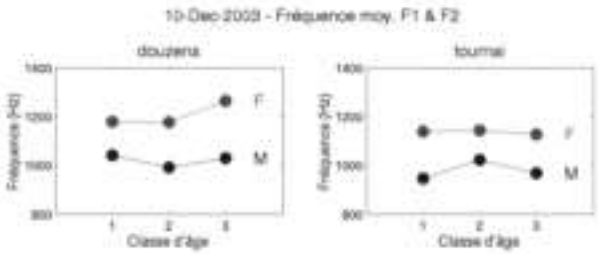

Figure 5 : Fréquence moyenne de F1 et de F2 en fonction de l'âge et du sexe pour les deux points d'enquête.

Différentes procédures de normalisation ont été mises au point dans le but de minimiser les variations liées aux différences anatomiques entre locuteurs (voir Disner, 1980 ; Rosner & Pickering, 1994 ; Syrdal & Gopal, 1986). Dans ce travail, nous avons entrepris d'évaluer trois de ces procédures, proposées la première par Gertsman (1968), la deuxième par Lobanov (1971), et la troisième par Nearey (1977).

La procédure de Gertsman est la suivante :

F_norm(i,j) = (F(i,j) – F_min(j)) / (F_max(j) – F_min(j))F(i,j) désigne la fréquence du j-ième formant pour la i-ème voyelle chez un locuteur donné, F_min(j) et F_max(j) une valeur minimale et une valeur maximale préétablies pour le j-ième formant. Cette procédure revient à rééchelonner les valeurs de fréquence de telle sorte qu'elles soient comprises entre deux bornes qui soient les mêmes pour tous les locuteurs, pour chaque formant.

La procédure de Lobanov est définie de la manière suivante :

F_norm(i,j) = (F(i,j) – mean(F(:,j))) / std(F(:,j))

Elle consiste à rapporter chaque valeur de fréquence à sa moyenne et à son écart type, de telle sorte que la moyenne et l'écart type soient les mêmes pour tous les locuteurs, pour chaque formant.

Enfin, la procédure de Nearey se définit comme suit :

F_norm(i,j) = log(F(i,j)) – mean(log(F(:,:)))Elle fait appel à une transformation non-linéaire (logarithmique) et à la moyenne de tous les formants pour chaque locuteur.

La qualité des différentes procédures de normalisation a été évaluée en calculant le rapport entre la dispersion intra-catégorie et la dispersion inter-catégories. Nous sommes partis du principe qu'une « bonne » procédure de normalisation tend à minimiser ce rapport, c'est-à-dire à réduire les variations observées pour chaque voyelle tout en préservant les différences entre voyelles. Les évaluations ont été restreintes aux voyelles dont le timbre est a priori bien établi pour les deux points d'enquête étudiés, ex. : *patte*, *épée*, *liège*, *creux*, *peuple*, etc. La figure 6 illustre les résultats obtenus au moyen des différentes procédures pour les données de Douzens.

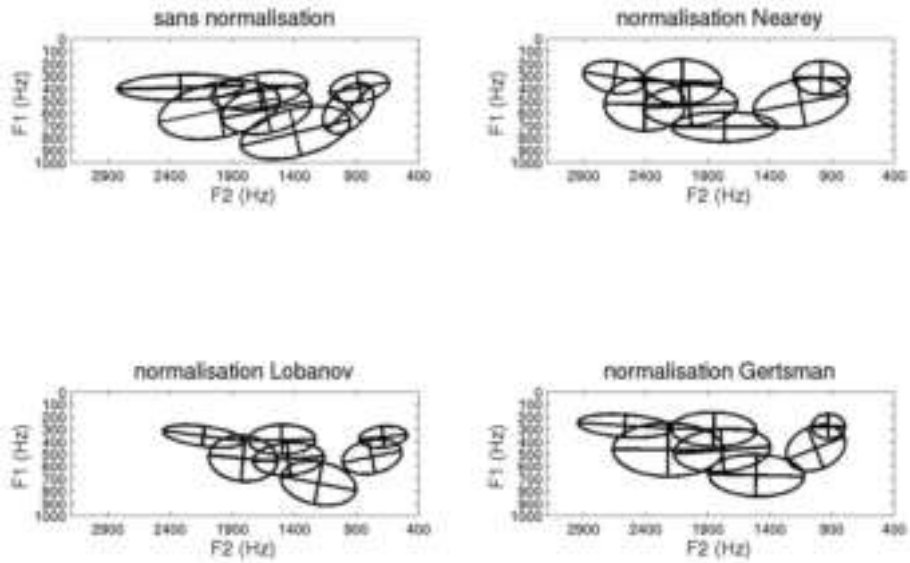

Figure 6 : Résultats des différentes procédures de normalisation. Données de Douzens.

La normalisation exerce un effet dans le bon sens lorsque la taille des ellipses diminue et que les ellipses sont plus éloignées les unes des autres, par comparaison avec les données initiales (sans normalisation). On constate que cet effet se manifeste à peu près dans les mêmes proportions pour les trois procédures testées. La figure 7 illustre les résultats obtenus pour les données de Tournai. Dans ce second cas, la procédure de Lobanov semble donner de meilleurs résultats que les deux autres.

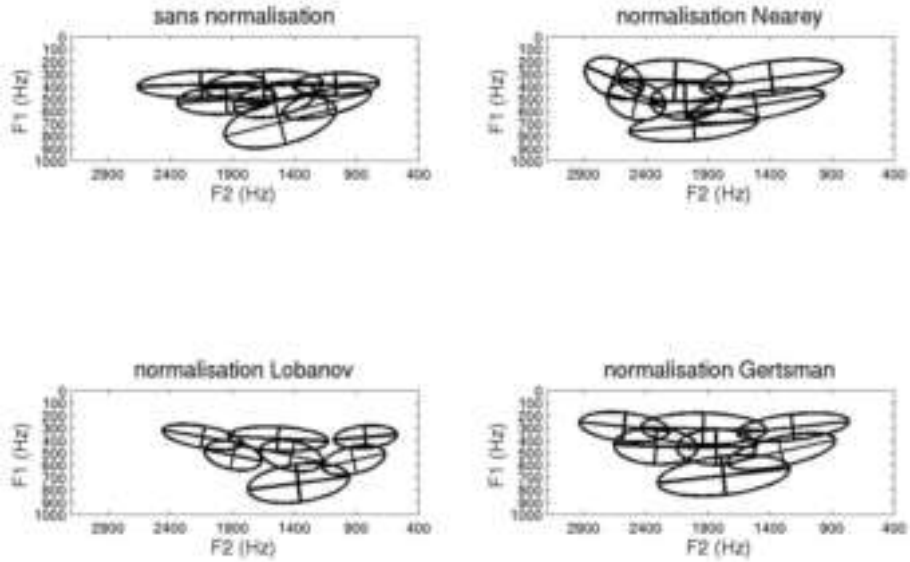

Figure 7 : Résultats des différentes procédures de normalisation. Données de Tournai.

Le rapport entre la dispersion intra-catégorie et la dispersion inter-catégories est présenté dans le tableau 6 pour les différentes procédures. Ce rapport est exprimé ici sous la forme d'un pourcentage. La valeur de ce rapport pour les données brutes est également indiquée à titre de comparaison.

	Douzens	Tournai
Données brutes	42 %	56 %
Nearey	30 %	46 %
Lobanov	29 %	34 %
Gertsman	31 %	45 %

Tableau 6 : Rapport intra/inter pour les différentes procédures de normalisation

Ces données confirment que les trois procédures aboutissent à des résultats équivalents pour Douzens, alors que la procédure de Lobanov se montre supérieure aux deux autres pour Tournai.

Sur les deux points d'enquête, la procédure de Lobanov est donc globalement la plus satisfaisante. Cependant, l'efficacité d'une procédure de normalisation varie d'un système phonologique à l'autre. La solution la meilleure semble donc de déterminer empiriquement, et pour chaque point d'enquête, la procédure de normalisation la mieux adaptée à la variété de français étudiée.

Remerciements

Ce travail est réalisé avec le soutien du projet PFC, dont nous remercions les responsables, Jacques Durand, Chantal Lyche et Bernard Laks. Nous remercions également Philippe Hambye et Géraldine Mallet pour avoir mis à notre disposition les enregistrements réalisés à Tournai et en Vendée, respectivement. Nous remercions enfin Dominique Fohr et Yves Laprie pour nous avoir donné accès à leur aligneur.

Références

Disner, S.F. (1980). Evaluation of vowel normalization procedures, *Journal of the Acoustical Society of America* 67, 253-261.

Gertsman, L.H. (1968). Classification of self-normalized vowels, *IEEE Trans. Audio Electroacoust.* AU-16, 78-80.

Labov, W. (2001). *Principles of Linguistic Change, vol. 2: Social Factors* (Blackwell, Malden, MA).

Lobanov, B.M. (1971). Classification of Russian vowels spoken by different speakers, *Journal of the Acoustical Society of America* 49, 606-608.

Nearey, T. (1977). *Phonetic Feature Systems for Vowels*, PhD Diss., University of Connecticut, Storrs, CT.

Nguyen, N., & Fagyal, Z. (2003). Acoustic aspects of vowel harmony in French, *XVth International Congress of Phonetic Sciences*, Barcelone, Espagne, 3-9 août 2003, pp. 3029-3032.

Rosner, B.S., & Pickering, J.B. *Vowel Perception and Production* (Oxford Univ. Press, Oxford, UK).

Syrdal, A.K., & Gopal, H.S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels, *Journal of the Acoustical Society of America* 79, 1086-1100.