Voyelles brèves en parole conversationnelle

Meunier C., Meynadier Y., Espesser R.

Laboratoire Parole & Langage – CNRS UMR 6057
Université de Provence, Aix-en-Provence, France
Christine.Meunier@lpl-aix.fr
http://www.lpl.univ-aix.fr/

ABSTRACT

This work deals with the phenomenon of vowel reduction in spontaneous speech. Automatic and manual analyses have been conducted on a large conversational speech corpus (CID) to study the extra-short vowels (less than 30 ms), generally excluded from the automatic analyses. A strong reduction in the vocalic system and very short durations are observed for a great proportion of vowels in the corpus. A manual analysis highlights the specific realisations of these extra-short vowels: they are more often in function words than in content words; almost all of them belong to monosyllabic words; their formant values show strong dispersion in F1/F2 plan; a large context is needed for their identification.
Keywords: vowels, spontaneous speech, reduction.

1. INTRODUCTION

Les travaux sur l'analyse de la parole s'intéressent de plus en plus aux caractéristiques de la parole non contrôlées. Il s'est avéré peu à peu essentiel de pouvoir donner une interprétation des formes de réalisation des sons dans leur contexte le plus courant, autrement dit la parole continue, spontanée, conversationnelle, etc., de façon à les mettre en perspective avec d'autres paramètres linguistiques [1]. Cette orientation permettrait de donner aux variations de la parole une interprétation plus globale et plus intégrée prenant en compte le langage dans sa globalité.

Toutefois, l'analyse phonétique de très larges corpus nécessite un travail de traitement considérable impossible à réaliser manuellement. Le recours à des analyses automatiques (alignement de la transcription sur le signal, phonétisation, analyses acoustiques automatisées, etc.) devient donc indispensable. Cependant, l'analyse automatique, outre les erreurs qu'elle engendre, écarte également nombre d'observations et données qui peuvent réduire, voire biaiser, l'interprétation des phénomènes considérés. Sans chercher à affaiblir les apports indéniables de l'automatisation des analyses (dont quelques descriptions seront présentées en premier lieu dans ce travail), nous proposons ici une observation pilote empirique et manuelle de données communément écartées par les analyses automatiques dans l'étude de la réduction vocalique en parole conversationnelle.

2. LE CORPUS CID ET SON EXPLOITATION

Nos analyses s'appuient sur l'exploitation des données phonétiques du *Corpus of Interactional Data* (CID) [2] qui constitue une ressource unique en son genre pour l'analyse du français parlée en interaction à différents niveaux : phonétique, prosodique, syntaxique, sémantique, pragmatique et mimo-gestuel. Il compte 8 dialogues d'environ 1 heure. Les 16 locuteurs (10 femmes et 6 hommes) d'origine régionale différente résident pour la majorité depuis plusieurs années dans la région du Sud-Est de la France. Les enregistrements ont été faits en chambre sourde et en stéréo (une piste par interlocuteur).

Une transcription orthographique enrichie des enregistrements a été effectuée par deux experts à l'aide du logiciel PRAAT. Puis, le corpus a été phonétisé. Les phonèmes ont enfin été alignés sur le signal à l'aide de l'aligneur du LORIA-INRIA [3] avec une résolution de 8 ms et une durée minimale de 24 ms.

2.1. Analyse automatique des voyelles

Sept catégories de voyelles orales sont analysées dans ce corpus : i, y, u, e ([e, ɛ]), @ ([ø, œ, ə]), o ([o, ɔ]) et A ([a, ɑ]). Les voyelles moyennes et ouvertes ne sont donc pas distinguées.

Les trois premiers formants sont automatiquement détectés au centre de la voyelle (ESPS package, Entropic, 1997) et calculés par LPC (méthode d'autocorrélation) avec une fenêtre de 49 ms (en cosinus^4). Afin d'éliminer les détections aberrantes, seules les voyelles comprises entre 30 et 300 ms ont été prises en compte dans l'extraction automatique de formants. Or, si les voyelles au delà de 300 ms représentent une proportion marginale de l'échantillon de voyelles (2% de l'effectif global), les voyelles inférieures à 30 ms représentent 30% de l'effectif. La proportion importante de ces voyelles extra-brèves mérite d'y porter attention, ce qui nécessite un repérage et une investigation manuels.

2.2. Analyse manuelle des voyelles

Notre attention s'est portée sur une partie significative des voyelles extra brèves (moins de 30 ms) de façon à évaluer la nature de ce type plutôt fréquent de réalisation. Nous avons pensé d'emblée que ces durées très courtes

pouvaient être dues en partie à des erreurs d'alignement. Deux experts ont donc corrigé manuellement l'alignement automatique sur deux locuteurs du corpus (AG : homme, et ML : femme). Leurs corrections font ressortir que l'alignement automatique sous-estime globalement de 20 ms la durée de ces voyelles. Notre analyse concerne ainsi un échantillon des cas de voyelles les plus brèves et correctement alignées, c'est-à-dire toujours inférieures à 30 ms (76 pour AG et 31 pou ML).

3. LES VOYELLES EN PAROLE CONVERSATIONNELLE

Les voyelles orales représentent environ 39% des 272 166 phonèmes du CID. La voyelle e est la plus fréquente devant la voyelle a. o, y et u montrent la plus faible proportion. Ensemble les trois voyelles e, A et @ représentent 70% des voyelles orales du corpus (Figure 3). Cette répartition est globalement comparable à celle décrite dans la littérature [4].

La plus grande partie des voyelles analysées automatiquement dure entre 50 et 100 ms. Si l'on compare les valeurs F1/F2 en Hertz de nos analyses à des valeurs prototypiques mesurées dans un corpus de syllabes [5], on observe une importante réduction du système vocalique du français (Figure 1). Cette réduction repose sur une diminution de F1 pour les voyelles ouvertes et mi-ouvertes (A, e, @, o), une augmentation de F2 pour les voyelles d'arrière (u, o) et une diminution de F2 pour les voyelles d'avant (i, e). Le F1 des voyelles fermées varie très peu. La voyelle la plus stable est /y/. Ces résultats sont conformes aux travaux portant sur la réduction vocalique [6] [7].

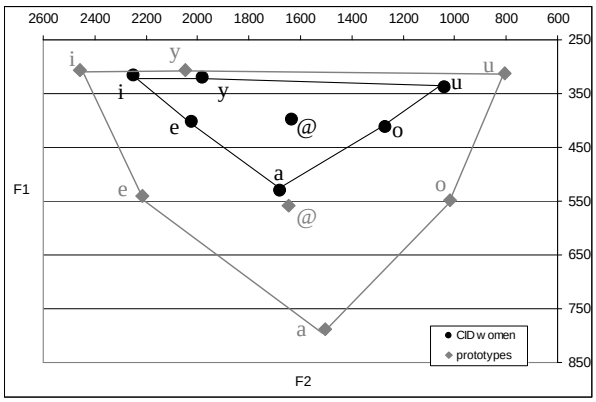

Figure 1 : Moyenne des formants des voyelles orales (femmes) dans le CID (en noir) et prototypiques (en gris)

La Figure 1 présente les valeurs des voyelles du CID sans tenir compte de leur durée. La Figure 2 montre un exemple d'évolution des valeurs de formants (F1 de A) en fonction de la durée des voyelles. Bien que la réduction formantique est très progressive et ne présente pas de seuil, la pente de cette réduction est sensiblement plus marquée pour les voyelles inférieures à 150 ms. Egalement sous 100 ms, on peut observer la grande dispersion des valeurs de formants, dont une partie peut être attribuée à des erreurs de détection. Pour les A les plus courts, la valeur moyenne du F1 est autour de 400 Hz, tandis que pour les A les plus longs la moyenne se situe vers 750 Hz, ce qui est proche des valeurs prototypiques.

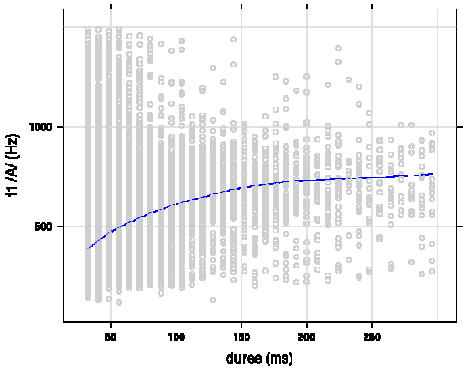

Figure 2 : Valeurs et tendance générale *(régression polynomiale locale, fonction loess de R)* du F1 de A en fonction de la durée

Les durées des voyelles en parole conversationnelle sont nettement plus courtes que les durées habituellement observées en parole lue. Mais plus précisément, une très forte concentration des données est observée sur des durées très courtes. Ainsi, après alignement sans aucune correction, 60% des voyelles durent moins de 40 ms.

4. VOYELLES EXTRA BRÈVES

L'étude de ces voyelles porte sur différents aspects en lien supposé avec leur durée brève : (a) la catégorie lexicale des mots, (b) la longueur syllabique des mots, (c) le dévoisement contextuel, (d) le timbre, selon son niveau d'identification et sa structure formantique (F1/F2) des voyelles extra-brèves (dorénavant EB). L'échantillon analysé manuellement de ces voyelles reflète la distribution des voyelles du CID (Figure 3). Ainsi, on retrouve les deux voyelles e et A les plus fréquentes et la moins fréquente u. Seul y est véritablement surreprésenté.

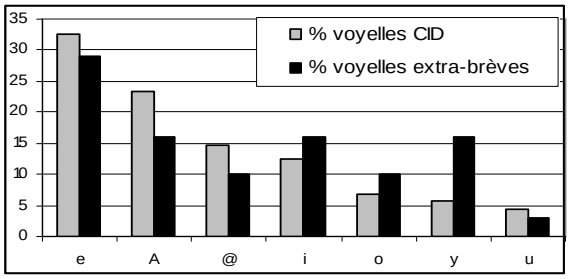

Figure 3 : Proportion des voyelles dans la totalité du CID et dans l'échantillon des voyelles extra brèves

4.1. Catégorie lexicale et longueur syllabique

Les voyelles EB sont annotées selon la catégorie lexicale et la longueur, de 1 à 5 syllabes du mot les contenant. Deux types de mots sont distingués : les mots grammaticaux (Gram) regroupant les pronoms, les prépositions, les conjonctions, les interjections et les verbes auxiliaires et semi-auxiliaires, et, les mots lexicaux (Lex) comprenant les noms communs, les adjectifs, les

adverbes et les verbes pleins.

Quel que soit le locuteur, plus de 75% de ces voyelles EB appartiennent à des mots grammaticaux maximalement bisyllabiques, et très majoritairement monosyllabiques : 54% pour AG et 65% pour ML (Figure 4). Le quart restant de ces voyelles apparaît dans des mots lexicaux comptant de 1 à 5 syllabes, avec une préférence pour les mots lexicaux courts (mono- ou bisyllabiques).

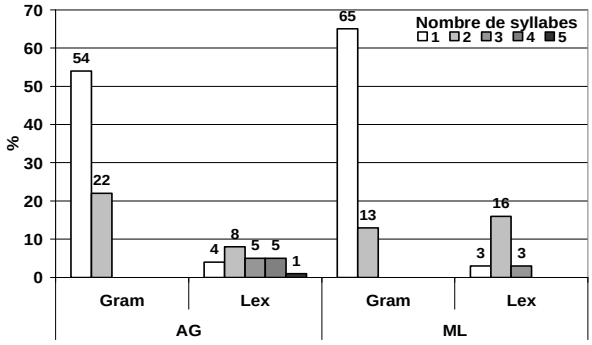

Figure 4 : Proportion des voyelles extra brèves selon la catégorie lexicale et la longueur syllabique du mot.

4.2. Dévoisement

Les cas de dévoisement, total ou partiel, des voyelles extra brèves ont également été relevés. La distribution des voyelles EB dévoisées (Table 1) a été observée selon les contextes phonétiques définis par la nature du segment précédant et suivant la voyelle, à savoir une consonne sourde (Cs), un segment vocalique ou consonantique voisé (Sv) ou une pause silencieuse (#). Les voyelles dévoisées représentent une portion non négligeable des voyelles EB: 22% chez AG et 16% chez ML. Les voyelles hautes, et plus nettement les antérieures i et y (particulièrement représentée par le pronom "tu"), sont les plus fréquemment dévoisées, même si toutes semblent pouvoir l'être. Les contextes de dévoisement vocalique impliquent nécessairement la contiguïté d'une consonne sourde, aucun cas n'étant constaté entre 2 segments voisés. Malgré une relative disparité entre les locuteurs (AG étant plus sensible au dévoisement que ML), la position d'une voyelle entre 2 consonnes sourdes est le contexte le plus favorable à un dévoisement fréquent : 63% des voyelles dans cette position pour AG et 33% pour ML. Aussi, seule une consonne sourde pré-vocalique implique plus nettement un dévoisement (un tiers des cas chez les 2 locuteurs) qu'une consonne sourde post-vocalique, dont l'influence est plus anecdotique.

4.3. Niveau d'identification et structure formantique

Le timbre des voyelles EB est étudié sous l'angle acoustique de leur structure formantique F1/F2 et sous l'angle perceptif du niveau de contexte nécessaire à leur identification. Le niveau contextuel d'identification a été déterminé empiriquement sur la base d'une écoute attentive des experts ayant corrigé l'alignement automatique. Pour chaque voyelle, ils ont écouté d'abord la voyelle isolément (V), puis la syllabe et/ou le mot (S/M) la contenant, puis le groupe de mots (L). Le niveau à partir duquel les experts ont eu la sensation de reconnaître la voyelle attendue détermine son niveau contextuel d'identification.

L'analyse acoustique du timbre laisse apparaître un espace vocalique réduit encore plus resserré que celui représentant la moyenne des réalisations du CID (Figure 5, locuteur AG). Les voyelles occupent une position relativement conforme à celle d'un espace prototypique. Toutefois les positions dans l'espace font ressortir un net décalage global du système vers une plus grande fermeture, et vers une antériorisation des voyelles postérieures o et u (le faible effectif de ces deux voyelles, 5 cas chacune, est cependant à souligner). Notons également que ces positions sont le reflet de moyennes. La distribution de l'ensemble des voyelles montre une très grande dispersion des réalisations (Figure 5, points) laissant penser que, dans ce type de réalisation, le rapport distinctif entre les voyelles ne relèverait pas essentiellement de leur timbre.

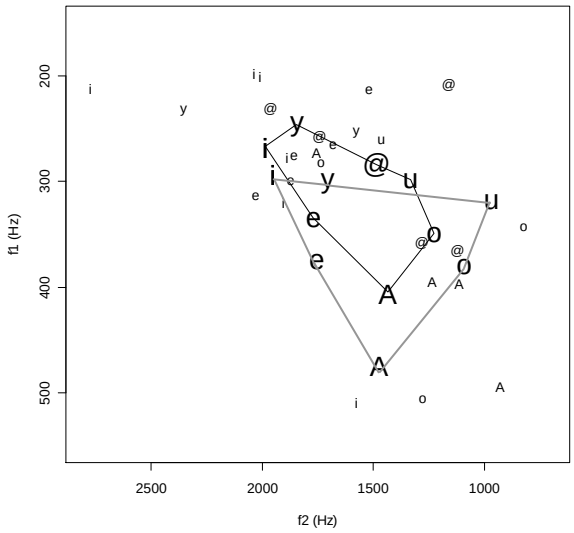

Figure 5 : Locuteur AG: moyennes des voyelles produites dans le CID *(trait gris)* et des voyelles EB *(trait noir)*. Les points isolés *(avec une police plus petite)* représentent les valeurs effectives de chaque voyelle EB.

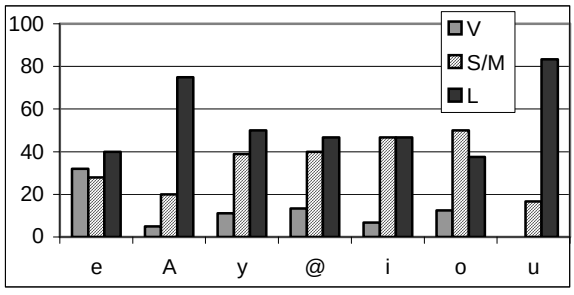

Figure 6: Taux d'identification des voyelles EB selon le niveau du contexte : V(oyelle), S(yllabe)/M(ot), L(arge).

Les voyelles EB sont majoritairement identifiées à un niveau contextuel large, à savoir supra lexical (V = 14% <

S/M = 34% < L = 52%). Les voyelles demandant le contexte le plus large pour être identifiées sont A et u, celle reconnue le plus fréquemment en contexte plus étroit est e (Figure 6). L'influence du type de mot sur l'identification des voyelles EB est également notable. Ainsi, les mots monosyllabiques grammaticaux (67%) impliquent bien plus fréquemment un contexte large que les lexicaux (25%). En outre, quelque soit le mot, la meilleure identification au niveau de la syllabe ou du mot concerne les mots bisyllabiques. Il est également remarqué que la longueur du mot ne semble pas particulièrement favoriser une meilleure reconnaissance des voyelles EB. Enfin, le dévoisement implique une identification plus difficile des voyelles EB. Ainsi, 64% des voyelles dévoisées sont identifiées au niveau large contre 47% pour les voyelles voisées. De même, 4,5% et 16,5% des voyelles respectivement dévoisées et voisées sont reconnues isolément.

5. CONCLUSIONS

Ce travail nous a permis de relever des tendances générales concernant les aspects quantitatifs et qualitatifs des réalisations des voyelles en parole conversationnelle. La réduction de l'espace vocalique observé reste cohérente avec les études menées dans ce domaine sur le français. Elle confirme une forte corrélation entre la diminution des durées et la réduction du système, indiquant que la plupart des voyelles réalisées en parole conversationnelle sont produites dans un espace très réduit. Cette étude fait ressortir, en outre, une grande proportion de voyelles très brèves où l'inventaire complet des voyelles est représenté. L'analyse de ces voyelles inférieures à 30 ms, communément écartées par les approches automatiques, semble cependant indiquer que le processus de réduction de l'espace vocalique en parole conversationnelle est quel que peu asymétrique, d'autres mécanismes pouvant intervenir lors de la production de voyelles très brèves.

Notre étude apporte également des observations sur la nature, l'occurrence et la perception de ces voyelles les plus brèves. Ces voyelles nécessitent un environnement contextuel le plus souvent plus large que le mot pour leur identification. On les retrouve très majoritairement dans des mots grammaticaux courts. Il semble ainsi que les mots grammaticaux soient surreprésentés dans les voyelles extra brèves. Nous avons en effet noté que dans le corpus CID, la proportion de mots lexicaux (48%) est légèrement plus importante que celle des mots grammaticaux (43%), alors que les voyelles EB sont 3 fois fréquentes dans des mots grammaticaux que lexicaux. Enfin, ces voyelles montrent une fréquence importante de dévoisement en voisinage de consonnes sourdes. Il serait intéressant de comparer ce fait au regard des voyelles plus longues, en supposant que les voyelles brèves sont plus soumises à l'influence phonétique contextuelle.

Les analyses automatiques permettent de dégager des informations phonétiques qu'il est devenu très difficile de valider de façon fiable par une inspection manuelle. Notamment, l'explication des phénomènes de variation, très importants dans ce type de corpus, n'est véritablement possible que par l'analyse d'une masse de données considérable assurant une interprétation plus globale intégrant de multiplies niveaux et facteurs linguistiques. Toutefois, en parallèle, l'investigation manuelle apporte une précision explicative pertinente des processus que des analyses automatiques tendent à sous-estimer.

***Remerciements**: à l'Institut de Linguistique Française (ILF) pour son soutien financier, à S. Clairet et à A. Coquillon.*

Table 1 : Distribution des voyelles EB dévoisées (D) et voisées (V) selon le contexte phonétique.

	AG									ML									
	Cs_Cs		Cs_Sv		Sv_Cs		Sv_Sv	total		Cs_Cs		Cs_Sv		Sv_Cs	Sv_Sv	#_Cs	#_Sv	total	
	D	V	D	V	D	V	D	D	V	D	V	D	V	V	V	V	V	D	V
e	2	2		1		5	5	**2**	13		2		3	2			1		8
A	1	1	1	1		1	9	1	12		1		1		2				4
y	2	1	3	1	1	1	4	**6**	7	1		1		1	2			**2**	3
@	1	1		4		2	4	1	11		1		1	1					3
i	3		1	2		2	2	**4**	6			2	1	1	1			**2**	3
o			1	3			1	1	4	1			1			1		1	2
u	1	1			1	1	1	**2**	3						1				1
nombre	10	6	6	12	2	12	26	17	56	2	4	3	7	5	6	1	1	5	24
% par contexte	63	38	32	63	14	86	92,9	22	74	33	67	30	70	100	85,7	100	33,3	16	75

BIBLIOGRAPHIE

[1] Johnson, K. Massive reduction in conversational American English, In K. Yoneyama & K. Maekawa (eds.) *Spontaneous Speech: Data and Analysis. Proceedings of the 1st Session of the 10th International Symposium.* Tokyo, Japan, 29-54, 2004.

[2] Bertrand, R et al. Le CID - Corpus of Interactional Data: protocoles, conventions, annotations, *TIPA*, 25, 25-55, 2007.

[3] Fohr, D., O. Mella, C. Cerisara, and I. Illina:, The automatic news transcription system: ANTS, some real time experiments, in *INTERSPEECH*, 377–380, 2004.

[4] New B., Pallier C., Ferrand L., Matos R. Une base de données lexicales du français contemporain sur internet: LEXIQUE, *L'Année Psychologique*, 101, 447-462, 2001. http://www.lexique.org

[5] Tubach J.P., Description acoustique, *La parole et son traitement automatique*, Masson, 79-130, 1989.

[6] Lindblom B., Spectrographic study of vowel reduction, *JASA*, 35, 1773-1781, 1963.

[7] Gendrot, C. & Adda, M. (2006) Is there a universal impact of duration on formant frequency values of oral vowels? An automated analysis of speech from eight languages. *Laboratory Phonology X*, 2006.