Comparaison de parole journalistique et de parole spontanée : analyses de séquences entre pauses

Cedric Gendrot[1], *Martine Adda-decker* [1,2] *et Carolin Schmid* [1,3]
(1) Laboratoire de Phonétique et Phonologie, UMR 7018 CNRS/Université Sorbonne Nouvelle,
(2) LIMSI, UPR 3251, bât. 508, rue John von Neumann, 91403, Orsay
(3) Université de Trier, FBII-Phonetik, 65296 Trier, Deutschland
cgendrot@univ-paris3.fr, schm2801@uni-trier.de, madda@limsi.fr

RESUME

Nous comparons le corpus de parole journalistique ESTER (Galliano et al., 2005) au corpus de parole spontanée NCCF (Nijmegen Corpus of Casual French, Torreira et al, 2010) en termes de durée, de f0 et de caractéristiques spectrales au sein de séquences suitées entre 2 pauses. Les montées de continuation de f0 sont en moyenne absentes pour la parole spontanée avec une ligne de déclinaison moins marquée. Pour ces 2 corpus, nous observons un allongement qui commence à partir de 60% de la durée de la séquence, mais significativement moins net en parole spontané. L'allongement de début de séquence est observé en parole journalistique seulement. Comme attendu, nous observons un débit plus important en parole spontanée avec des durées de phonèmes plus courtes impliquant une réduction vocalique plus importante.

ABSTRACT

Comparison of journalistic and spontaneous speech: analysis of sequences between pauses.

In this study we compare the ESTER corpus of journalistic speech (Galliano et al., 2005) and the NCCF corpus of spontaneous speech (Torreira et al, 2010) in terms of duration, f0 and spectral reduction in productions automatically detected as sequences between, pauses. Continuation f0 rises are overall absent in spontaneous speech and sequences reveal a declination slope with less amplitude than in journalistic speech. For both corpora, lengthening starts around 60% of the sequence duration, but significantly less in spontaneous speech. Lengthening in the initial part of the sequence is observed in journalistic speech only. As expected we measure a faster speech rate in spontaneous speech with shorter vowel durations implying a more important vowel reduction.

MOTS-CLES : parole journalistique, spontané, déclinaison, f0, durée, réduction spectrale.
KEYWORDS : journalistic speech, spontaneous, declination line, f0, duration, spectral reduction.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 649–656,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Depuis quelques années, avec l'amélioration des systèmes de transcription et d'alignement automatique de la parole, l'accès aux corpus de parole continue (préparée dans un premier temps, puis spontanée plus récemment) est devenu possible. Jusqu'alors, les analyses effectuées sur la parole spontanée ne concernaient que quelques dizaines de minutes ou quelques heures au maximum. Nous présenterons les résultats d'une étude préliminaire comparant un corpus de parole journalistique et un corpus de parole spontanée, d'environ 30 heures chacun. L'analyse de la parole spontanée trouve un intérêt particulier puisqu'étant la plus représentative possible d'une situation naturelle, elle permet de préciser les modèles de production et de perception de la parole, par exemple les modèles phonologiques (Ernestus et Baayen, 2011), les modèles prosodiques (Post, 1993) et les modèles de perception à exemplaires (Ernestus et Baayen, 2011).

A l'instar de Schmid et al. (accepté), nos analyses porteront principalement sur l'analyse de séquences situées entre pauses. Ces analyses permettront de modéliser les variations prosodiques de durée et de f0 des phonèmes en fonction de leur position dans ces séquences. Ces variations ayant un impact sur la réalisation des phonèmes (Lindblom, 1990 ; Gendrot et Adda-Decker, 2006), nous analyserons également les phénomènes de réduction sur ces données. Nous utiliserons pour le français journalistique le corpus ESTER dont le détail a été mentionné dans Galliano et al. (2005). Le corpus de parole journalistique est considéré comme de la parole préparée plutôt que lue, avec quelques séquences de parole libre. Le corpus NCCF (*Nijmegen Corpus of Casual French*), détaillé dans Torreira et al. (2010), a été utilisé pour représenter la parole spontanée et sera comparé au corpus ESTER. Dans les 2 cas, la segmentation et transcription orthographique a été dans un premier temps effectuée par des auditeurs humains et l'alignement en phonèmes et en mots a été réalisée automatiquement par le système d'alignement automatique du LIMSI (Gauvain et al. 2002).

En comparaison avec les corpus crées ad-hoc, l'utilisation de grands corpus (« phonétique de corpus ») présente l'avantage de la quantité mais présente malheureusement quelques inconvénients. En dehors du respect de la distribution des catégories, il peut également exister des problèmes de sélection des unités à analyser, comme c'est le cas par exemple pour les unités prosodiques mentionnées dans la littérature et qui sont fréquemment relevées d'après des approches phonologiques (*Top-Down)*. Notre travail porte ici sur des séquences de parole qui se rapprochent des groupes intonatifs (Jun et Fougeron, 2000). Une annotation manuelle étant difficilement envisageable sur des corpus dont la durée totale excède les 60 heures, il est nécessaire d'automatiser la procédure de détection de ces unités. Nous avons effectué cette détection en considérant comme séquences d'analyse les productions situées entre 2 pauses détectées comme silences de plus de 500ms par le système d'alignement automatique (cf. table 1). Toute séquence contenant une pause de plus de 50ms fut exclue des analyses ultérieures.

A l'intérieur de ces séquences, nous avons effectué des mesures de durée des phonèmes et mots tels que segmentés automatiquement par l'alignement automatique. Les mesures de f0 et les mesures spectrales sont quant à elles effectuées sur les parties centrales des voyelles au moyen de PRAAT, les précautions d'usage sont détaillées dans Gendrot et

Adda-Decker 2005, ainsi que Schmid et al. (accepté).

1.1 Hypothèses : questions de travail

L'analyse préliminaire que nous présentons a été effectuée depuis un découpage en séquences, dont le critère de sélection a été la présence de pauses telles que détectées par le système d'alignement son-texte. Il s'agit d'analyser à l'intérieur de ces séquences les mouvements de f0 (incluant les phénomènes de déclinaison) ainsi que les phénomènes d'allongements et de réduction.

	nombre	Durée moyenne en (s)	Ecart-type	Débit moyen (phon/s)
p. journalistique	562935	2.71	1.43	13.6
p. spontanée	516933	1.68	1.15	15.3

TABLE 1 – Caractéristiques principales des séquences dans les 2 corpus.

Nous ne comparons pas ici de la parole spontanée à de la parole lue, mais à un corpus de parole journalistique. La parole journalistique peut être qualifiée de style à part entière puisqu'elle implique une quantité plus importante d'accents lexicaux initiaux (Vaissière, 1997). En terme de débit, elle pourrait se situer à mi-chemin entre la parole lue et la parole spontanée. La parole journalistique peut être qualifiée de parole publique : l'articulation, sans être soutenue, y reste bonne, afin que la parole puisse être partagée par une large audience : on observe peu d'hésitations, peu de fragments de mots et les structures syntaxiques restent souvent proches du langage écrit. Les phénomènes de réduction y sont sans doute moindres que dans une vraie parole spontanée. Nous avons l'occasion de quantifier cette prédiction ici.

D'après des études réalisées précédemment sur de la parole journalistique, nous avions observé des phénomènes de réduction en fonction de la durée des voyelles et ce, bien que le français ne soit pas une langue à accent lexical. Dans un premier temps, nous chercherons à savoir si la réduction observée pour de la parole journalistique peut encore être accrue pour de la parole spontanée. Le cas échéant, serait-elle due uniquement à des différences de durée phonémique (et par extension de débit), ou bien à des durées comparables, pourrait-on observer des différences de réduction ?

Nous effectuerons également des mesures de f0 permettant de calculer la ligne de déclinaison. Celle-ci est définie comme la tendance de la fréquence fondamentale de baisser au cours de la phrase (T'Hart et al., 1990), entre une ligne supérieure qui relie ses sommets et une ligne inférieure qui relie ses vallées qui descendent toutes deux également. Certaines caractéristiques de la ligne de déclinaison restent malgré tout discutées comme par exemple ses variations inter-langues ou son aspect conscient chez le locuteur. Selon un protocole semblable à celui utilisé par Yuan et Liberman (2010) et par Schmid et al. (accepté) qui ont comparé la ligne de déclinaison de plusieurs langues, nous comparons ici la ligne de déclinaison pour la parole journalistique et la parole spontanée. Ces résultats pourraient permettre de comprendre si la ligne de déclinaison

est programmée en partie, voire en totalité par le locuteur. En effet, en parole préparée, le locuteur a une idée plus précise de la longueur totale de la phrase dès le début de sa production, ce qui n'est pas nécessairement le cas en parole spontanée, notamment pour des séquences d'une durée assez longue (au-delà de 3 secondes).

Les variations de durée seront également analysées avec intérêt : en effet, les modèles prosodiques ne prennent que peu en compte les variations de durée, principalement à cause des variations importantes de durée intrinsèque des différents phonèmes. Or il pourrait apparaître que les phénomènes d'allongement caractérisant les fins d'unités prosodiques de haut niveau (comme le groupe accentuel ou le groupe intonatif de (Jun et Fougeron, 2000) sont prédominants par rapport aux phénomènes mélodiques, et ce particulièrement dans le cas de la parole spontanée. Pour ce faire, nous avons mesuré les durées des phonèmes en contexte suivant (et non suivant et précédent afin de préserver un nombre important d'occurrences) sur l'ensemble du corpus NCCF. Les durées ainsi calculées en termes de ratio par rapport à ces valeurs de référence qui sont disponibles ici : http://www.personnels.univ-paris3.fr/users/cgendrot/pub/download/durees_phonemes_en_contexte.txt

2 Comparaison des valeurs de durée

Comme mesuré par Nooteboom (1997) pour la longueur des phonèmes à l'intérieur des mots, parmi les séquences que nous avons analysées, plus la séquence mesurée est longue, plus le nombre de phonèmes contenus dans cette séquence est important, et plus la durée de ces phonèmes est faible (figure 1)

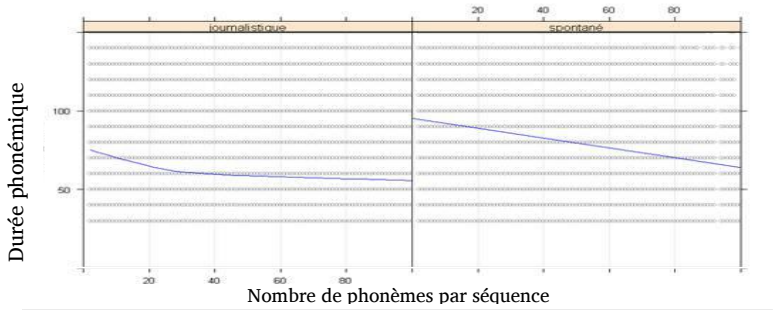

Figure 1 : mesure de durée phonémique en fonction du nombre de phonème dans la séquence. A gauche parole journalistique et à droite parole spontanée.

Pour la figure 2 ci-dessous, la durée de chaque phonème (normalisée par rapport aux valeurs de durée de référence) est affichée en fonction de sa position au sein de la séquence (en pourcentage de durée). Pour les 2 corpus, nous observons un allongement qui commence à partir de 60% de la durée de la séquence, mais significativement moins net en parole spontanée. L'allongement de début de séquence est observé en parole journalistique seulement. Ces résultats sont observés quelque soit la durée de la séquence.

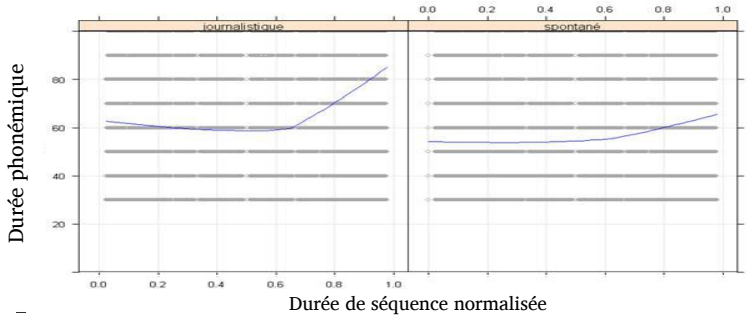

Figure 2 : mesure de durée phonémique normalisée en fonction de la position dans la séquence. A gauche parole journalistique et à droite parole spontanée.

3 Comparaison des valeurs de f0

D'après les procédures détaillées dans Schmid et al. (accepté) et Yuan et Liberman (2010), nous avons pu recueillir les contours de f0 lissés pour chacune des séquences et mesurer la pente de ce contour par une régression linéaire. Dans les 2 corpus, la pente moyenne est fortement dépendante de la longueur de la séquence comme le montre la figure 3 : plus la phrase est longue et plus la pente est mesurée comme faible, et ce particulièrement en parole spontanée. (corrélation de Pearson : $r^2 = 0.4$ en parole journalistique contre $r^2 = 0.31$ en parole spontanée). La valeur moyenne de la pente de la ligne de déclinaison est plus faible en parole spontanée (-2.48 demi-tons/seconde pour la parole journalistique contre 2.25 pour la parole spontanée, différence significative à $p < 0.0001$ pour un test-t).

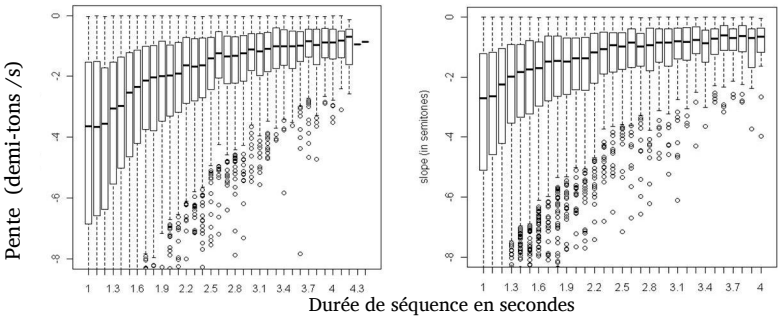

Figure 3 : pente moyenne (et écart-type) de la f0 en fonction de la durée de la séquence. A gauche parole journalistique et à droite parole spontanée.

Les figures présentées ci-dessous découpent le contour de f0 en une ligne supérieure (pics

de f0) et une ligne de base (vallées). Quelque soit la durée des séquences, nous pouvons observer que les montées de continuation sont faibles voire absentes pour le corpus de parole spontanée (figures 4 et 5). Pour les phrases inférieures à 2 secondes, les montées de f0 initiales dont le maximum se situe à environ 15% du début de la séquence sont présentes bien que moins amples en parole spontanée. En analysant les séquences de durée croissante (de 1 à 2 secondes, puis 2 à 3 secondes, etc), nous pouvons remarquer que la f0 (ligne supérieure et ligne de base) voit sa ligne de déclinaison relevée (plus plate) à partir de la moitié de la séquence pour les séquences de plus de 3 à 4 secondes (figures 4 et 5)..

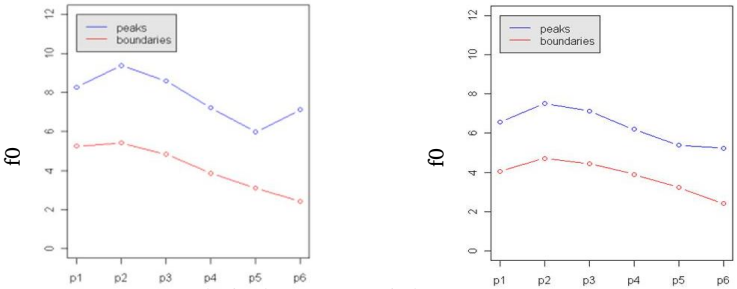

Figure 4 : contour de f0 normalisé en ligne supérieure et ligne de base. Temps normalisé. A gauche parole journalistique et à droite parole spontanée. Séquences de 2 secondes.

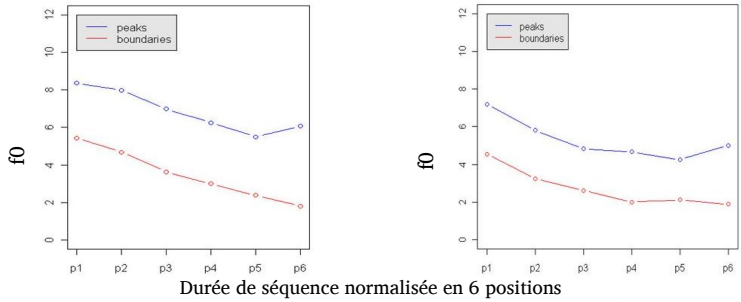

Figure 5 : contour de f0 normalisé en ligne supérieure et ligne de base. Temps normalisé. A gauche parole journalistique et à droite parole spontanée. Séquences de 5 secondes.

4 Résultats : réduction spectrale

Après avoir observé des valeurs de débit plus élevées, et des durées vocaliques plus courtes en parole spontanée, nous pouvons visualiser ci-dessous l'espace vocalique

fournissant un indice de la réduction vocalique ci-dessous. Nous pouvons constater que la réduction vocalique est plus importante (l'espace vocalique étant plus petit) en parole spontanée (figure 6 gauche). Cependant, en considérant des catégories de durée comparable (de 30 à 60 ms, figure 6 droite), aucune différence d'espace vocalique n'apparaît alors entre les 2 corpus. Le décalage sur le 1er formant (F1) pourrait être expliqué par des différences de f0 entre les 2 corpus et sera détaillé dans la suite de nos travaux.

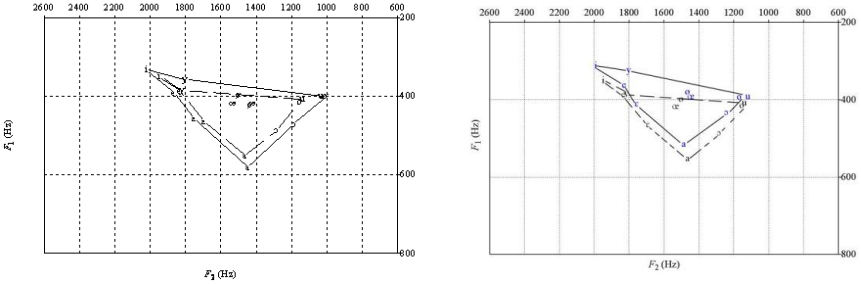

Figure 6 : Espace vocalique pour les locuteurs masculins en parole journalistique (traits pleins) vs. parole spontanée (pointillés). A gauche, toutes catégories de durée ; à droite, voyelles entre 30 et 60 ms.

5 Conclusion

Les études de corpus faites sur la parole spontanée permettent de mettre à jour des phénomènes décrits dans la littérature sur des données parfois peu importantes. Nous avons pu préciser ici certaines caractéristiques prosodiques dans le passage de la parole journalistique à la parole spontanée.

Les phénomènes de réduction, comme attendu, sont plus importants en parole spontanée et ils peuvent être prédits par le débit et/ou la durée des voyelles analysées. L'allongement final qui est conservé en parole spontanée, contrairement à la montée de continuation de f0, démarre à partir de 60% de la séquence.

Les analyses sur la ligne de déclinaison nous permettent de suggérer que les 2 lignes calculées permettent de distinguer une ligne de base liée à la physiologie, semblable entre les 2 styles de parole et une ligne supérieure plus dépendante du style. Nous émettons l'hypothèse que le planning des séquences étant moins prévisible en parole spontanée qu'en parole journalistique, pour les séquences plus longues (au-delà de 3 secondes) il est difficile pour le locuteur d'anticiper la ligne de déclinaison et nous observons un redressement de la ligne de déclinaison sur la 2ème moitié de la séquence.

Remerciements

Cette étude a été financée grâce au soutien de l'ANR ETAPE et Labex EFL.

Ernestus, M., & Baayen, R. H. (2011). Corpora and exemplars in phonology. In J. A. Goldsmith, J. Riggle, & A. C. Yu (Eds.), *The handbook of phonological theory (2nd ed.)* pages 374-400. Oxford: Wiley-Blackwell.

Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.-F. et Gravier, G., (2005). ESTER Phase II evaluation campaign for the rich transcription of French broadcast newshase II Evaluation campaignfor the rich transcription of French broadcast news. In: *Proceedings of Interspeech* 2005, pages 2453–2456.

GAUVAIN, J.L., LAMEL, L. et ADDA, G. (2002) The Limsi Broadcast News Transcription System, *Speech Communication*, 37(1-2): pages 89-108.

Gendrot, C. & Adda, M. (2005). Impact of duration on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German. *Proceedings of Eurospeech* – Lisbon (Portugal), September 2005, pages 2453-2456.

Gendrot, C. et Adda-Decker, M. (2006) Analyses formantiques automatiques en français : périphéralité des voyelles orales en fonction de la position prosodique. *26èmes Journées d'Etude de la Parole*, 12-16 juin 2006. pages 205-208.

Jun S.-A. & Fougeron C. (2000), A Phonological model of French intonation. In A. Botinis (ed.) *Intonation: Analysis, Modeling and Technology*. Dordrecht : Kluwer Academic Publishers. pages 209-242.

Lindblom B., 1990, Explaining phonetic variation : a sketch of the H & H theory, in *Speech production and speech modelling*, W. Hardcastle et A. Marchal, Dordrecht, Kluwer, pages 403-440

Nooteboom, S. G. (1997). The prosody of speech: Melody and rhythm. In W. J. Hardcastle & J. Laver (Eds.), *The Handbook of Phonetic Sciences*. Oxford: Blackwell. pages 640-673.

Post B., (1993), A phonological analysis of French intonation, *MA Thesis*, University of Nijmegen.

Schmid, C., Gendrot, C. et Adda-Decker, M. (accepté). F0 déclinaison: une comparaison entre le français et l'allemand journalistique. *29èmes Journée d'Etude de la Parole*, juin 2012, Grenoble.

T'Hart, Cohen et Collier (1990). *A perceptual study of intonation : An experimental-phonetic approach to speech melody*. Cambridge University Press.

Torreira, F., Adda-Decker, M., & Ernestus, M. (2010). The Nijmegen corpus of casual French. *Speech Communication*, 52, pages 201-212.

Vaissière, J. (1997). Ivan Fonagy et la notation prosodique. *Polyphonie pour Ivan Fonagy*. J. Perrot. Paris, L'Harmattan: pages 479-488.