Discrimination de styles de parole par analyse prosodique semi-automatique

Jean-Philippe Goldman[1,2], Antoine Auchlin[1], Anne Catherine Simon[2]

jean-philippe.goldman@unige.ch, antoine.auchlin@unige.ch,
anne-catherine.simon@uclouvain.be

[1]Département de Linguistique, Université de Genève, Suisse
[2]Institut Langage & Communication, UCLouvain, Belgique

Abstract

This study focuses on prosodic differences between speaking styles, and their automatic distinction. We aim at characterizing speaking styles with the purpose of distinguishing them from each other, modeling them and eventually adding expressivity to text-to-speech systems. This can be done with a multi-level annotation of varied corpora, based on automatic processes (like phonetic segmentation, grammatical tagging) and manual annotations (of perceived syllabic prominences and of delivery speech objects). Quantitative comparisons of various prosodic parameters are conducted through the acoustic and linguistic dimension to catch differences between speaking genres.

1. Introduction

Notre recherche porte sur la discrimination de phonostyles, c'est-à-dire de styles de parole perçus comme identifiant une situation de communication, via un genre, une image acoustique typifiée. Notre approche est basée sur l'examen de corpus. Notre objectif est de comprendre quels paramètres prosodiques discriminent certains genres de parole, afin de mieux décrire ces genres et d'implémenter, en synthèse, les styles correspondants.

Quels paramètres prosodiques permettent de discriminer des phonostyles, ces styles sonores caractéristiques d'un individu, d'un groupe social ou d'une circonstance de parole (Léon 1993: 3) ? Pour répondre à cette question, notre étude se base sur l'analyse d'une partie du corpus C-PROM (Avanzi et al. 2010) comprenant 6 genres de parole (lecture, conférence scientifique, interview radiophonique, journal parlé, récit conversationnel et discours politique) représentés chacun par trois échantillons. Notre objectif est de comprendre quels paramètres prosodiques discriminent le mieux une partie ou l'ensemble de ces genres de parole, afin de les décrire de manière appropriée et d'implémenter, en synthèse, les styles correspondants.

La définition comparée des styles de parole n'est pas une opération triviale: d'une part, les styles ne s'échelonnent pas aisément sur un seul axe (par ex. du plus au moins formel) permettant de faire des prédictions sur leurs caractéristiques prosodiques; d'autre part, les styles peuvent se décrire selon une matrice de traits situationnels partiellement indépendants du degré de formalité évoqué ci-dessus (par ex. parole publique vs. privée; monologue vs. dialogue) qui, on le verra, semblent être corrélés avec certains paramètres prosodiques. Cette étude (voir § 2) prend donc à la fois en compte chaque style-genre et les groupements de ces derniers selon des traits situationnels qui ont une influence sur le mode de conception du discours (Koch & Oesterreicher 2001).

Méthodologiquement, nous avons privilégié une description semi-automatisée de la prosodie (voir § 3), basée sur des outils libres d'accès et permettant d'obtenir pour chaque échantillon ou groupe d'échantillons une série de mesures pouvant être contrastées: débit,

registre et chemin de f0, proportion et localisation des syllabes proéminentes, etc. Les mesures, et les résultats, concernent tantôt les syllabes tantôt les phénomènes prosodiques se réalisant dans les macro–unités, définies ici comme « unités séparées par des pauses ».

2. Définir les styles-genres de parole

Commençons par une précision concernant l'emploi des termes *genre* et *style* ici. Johns-Lewis (1986) amalgame dans les 'discourse modes' (modes de discours) ce que Hymes désigne par « genre » (prière ; lecture ; poésie) et ce que Crystal & Davy nomment « modality » (bulletin de nouvelles radio ; commentaire sportif en direct…). Ces termes et ces exemples renvoient à une catégorisation préalable, situationnelle, de contraintes et déterminations donnant des formes particulières à la parole. Léon (1993, ch.8), après Lucci (1983), utilise également le terme de *genre* pour désigner cette catégorisation, et celui de *phonostyle* pour désigner les caractéristiques effectives d'une parole donnée – quoi que de manière non systématique ; le terme de *phonostyle*, chez Fónagy, Léon, entre autres, étant souvent également entendu comme hyperonyme. Nous distinguons ici, partout où c'est possible, le *genre*, classification basée sur le type d'activité de parole et le type de circonstances dans lesquelles elle est produite ; du *style*, ensemble de propriétés d'un échantillon (ou d'un groupe d'échantillons) de parole donné, qui appartient à un genre donné et le reflète plus ou moins. Ainsi, méthodologiquement, on étudie un *phonogenre* lorsqu'on regroupe des échantillons de parole selon leurs conditions de production (par genre) et qu'on étudie le profil prosodique commun (moyen) ; et l'on étudie un *phonostyle*, singulier ou prototypique, si l'on étudie les échantillons et qu'on les groupe selon leurs propriétés, communes ou distinctives.

Deux types d'approches sont illustrés dans les études prosodiques sur les styles de parole. Des auteurs comme Pierre Léon ou Ivan Fónagy se sont attachés à décrire des phonostyles typiques d'une manière qualitative, en faisant ressortir telle particularité (allongement de la syllabe pénultième de groupe intonatif, présence d'une certaine courbe mélodique, etc.) associée à tel type de locuteur / situation de parole. Des descriptions très fines et détaillées ont été produites[1], qui forment de bonnes sources d'hypothèses.

D'autre part, il existe une tradition d'études qui visent à comparer, souvent deux à deux et de manière partiellement automatisée, des échantillons de parole représentatifs de genres pour voir selon quels indicateurs prosodiques ils divergent de manière significative. On y apprend par exemple que

- du point de vue de la structure temporelle: le type et la distribution des pauses varie entre la lecture et la parole spontanée (Guaïtella 1997; Hirschberg 2000); l'allongement des voyelles sous l'accent en français varie entre la lecture (allongement des accents finaux plus marqué) et l'interview (accents initiaux plus marqués) (Astésano 1999); le débit de parole est plus élevé en parole lue que spontanée (Hirschberg 2000: 336 sur des données en anglais américain; cependant Koopmans & van Beinum 1991 font l'observation inverse sur du néerlandais);
- du point de vue tonal: le registre mélodique est plus compact en parole spontanée qu'en parole lue, et, pour la parole radiophonique, le registre mélodique est plus réduit pour les informations que pour les commentaires sportifs;
- du point de vue des unités intonatives: on compte plus d'unités intonatives mineures par unité intonative majeure en parole spontanée qu'en parole formelle, et la parole

[1] Par exemple, sur la manière de parler de Brigitte Bardot, sur l'accent du Midi, sur le style des journalistes à la radio, etc. (Fónagy 1983; Léon 1993 ; Callamand 1987).

formelle se caractérise par des unités intonatives majeures plus longues (Cid & Corugedo, cités par Llisterri 1992: 14; Degand & Simon 2009); par contre il n'a pas encore été démontré qu'il y ait des différences dans le choix des types de contours mélodiques selon le style (ou le genre) de parole (Hirschberg 2000: 345; Llisterri 1992).

Ainsi, on trouve dans la littérature d'une part des descriptions exhaustives d'un style de parole typé (un genre), et de l'autre des informations sur des paramètres prosodiques qui varient de manière systématique selon le style, la notion de style étant souvent réduite à une opposition entre deux styles (genres), typiquement entre parole lue et parole spontanée ou entre style formel et style informel.

Dès lors qu'on cherche à comparer plus de deux styles (ou genres) de parole, il devient difficile de les classer sur une unique échelle, de formalité, par exemple. Ainsi, pour prendre 4 genres parmi les 6 que compte de notre corpus - le journal parlé radiophonique, la lecture à haute voix (dans le cadre d'une tâche de recueil de données élicitées), la conférence scientifique et le discours politique - lequel devra être considéré comme le plus formel? De même, peut-on regrouper sous l'étiquette uniforme de parole lue la lecture à haute voix et le journal parlé radiophonique au titre que le locuteur lit un texte écrit à l'avance? Plusieurs aspects d'une situation de communication (discours préparé ou improvisé; discours public ou intime; discours monologal ou dialogal) se trouvent amalgamés si l'on tente de réduire la variation à un axe unique (formel vs informel), et des situations très différentes peuvent se trouver assimilées les unes aux autres.

Afin d'éviter ces écueils, nous faisons les propositions méthodologiques suivantes:

- identifier un genre de parole en précisant la tâche communicative spécifique accomplie, afin d'éviter que des tâches aussi différentes que des réponses produites dans le cadre de dialogue homme-machine (voir corpus étudié dans Hirschberg 2000) ou une narration conversationnelle entre amis soient catégorisées comme *parole spontanée* alors qu'elles diffèrent grandement entre elles;
- caractériser chaque genre selon les trois axes suivants (Llisterri 1992; Eskenazi 1993; Koch & Oesterreicher 2001; Simon et al. 2009[2]): (i) discours préparé (lu)[3] - improvisé (non lu); (ii) type d'audience (0=micro - face à face - beaucoup)[4]; (iii) discours médiatique (radio- ou télédiffusé) - discours non médiatique.
- essayer de relier différents paramètres de la variation prosodique (débit, registre tonal, pauses, etc.) à ces dimensions multiples qui caractérisent les genres de parole.

Nous discrétisons en traits des dimensions réputées graduelles pour des raisons essentiellement pratiques. De même l'inventaire des traits situationnels retenus est délibérément restreint (suffisant à discriminer nos situations de parole) et synthétique, chaque trait neutralisant différentes facettes[5]. Plusieurs avantages découlent de cette méthodologie:

- on peut à la fois opposer les genres de parole, mais aussi analyser leurs ressemblances en fonction des caractéristiques situationnelles qu'ils partagent (par ex. deux types de lecture - lire une histoire à un enfant vs lire un texte en laboratoire - partagent le trait « discours préparé » mais s'opposent sur celui du « type d'audience »);

[2] Une étude préalable (Simon et al. 2009) nous a permis de montrer que certains paramètres prosodiques (comme l'étendue du registre tonal ou le débit de parole) varient systématiquement selon certains aspects de la situation. Voir aussi (Lucci 1983, Campbell 2004).

[3] Le degré ultime de la parole préparée étant la parole lue.

[4] Ce trait renvoie au *cadre interactionnel* de Roulet et al. (2001), et plus précisément à la *réciprocité* (ou non), et aux conditions de *co-présence* spatiale et temporelle des interactants.

[5] Nous laissons ainsi de côté des traits importants, comme le caractère naturel - élicité (laboratoire) des données, ou le degré d'intelligibilité recherché (Eskenazy 1993: 502) et l'effort fourni par le locuteur.

- on peut mieux comparer les résultats des études publiées ou, tout au moins, statuer sur la comparabilité des résultats les uns avec les autres.

Cette méthodologie repose sur l'hypothèse, forte, d'un déterminisme des conditions de production de la parole sur ses propriétés, prosodiques et formelles. Cette détermination s'exerce de façon plus ou moins rigide, selon le degré de prototypicité de la situation de production, et se reflète dans l'homogénéité vs la dispersion de la variation prosodique - ce qui s'inscrit dans le cadre de notre hypothèse. Mais ceci définit les propriétés de *genre* (de parole), plutôt que de *style*. Le genre est un objet typifié, associé à un ensemble d'attentes normatives. Le style quant à lui doit être vu comme ce qui émerge de la parole dans un genre donné, et satisfait plus ou moins les attentes associées à ce genre. Le style, en retour, peut re-configurer le genre, et par là la situation dans laquelle il apparaît ou est supposé apparaître (Johns-Lewis 1986).

Tel est bien le défi que doit relever la synthèse de la parole: il faut que la parole synthétique présente les propriétés stylistiques typiques de genres de parole déterminés, permettant à l'auditeur de l'intégrer à tel ou tel mode, régime de production, dans telle ou telle situation - réelle ou virtuelle.

Aussi notre analyse ne repose-t-elle que temporairement sur l'hypothèse de la détermination situationnelle, hypothèse dont dépend par exemple la comparaison des mesures de dispersion par genre, tel genre apparaissant comme plus compact que tel autre. Les comparaisons faites sur les valeurs chiffrées de l'ensemble des syllabes en revanche s'inscrivent dans le cadre complémentaire, émergentiste: ce sont les propriétés des échantillons qui déterminent leurs assemblages en familles stylistiques. Elles visent à faire apparaître des ressemblances entre profils prosodiques, dans certaines dimensions; elles permettent également de proposer une hiérarchisation raisonnée de paramètres situationnels, en fonction de leur « impact stylistique », leur degré de marquage ou neutralisation dans les échantillons.

3. Méthodologie: analyse automatique de la prosodie

3.1. Présentation du corpus d'étude

Notre corpus comprend 17 échantillons de parole d'une durée moyenne de 210 secondes, appartenant à 6 genres ou conditions de production : (i) lecture à voix haute d'un texte (lecture d'un article de journal, situation neutre); (ii) journal parlé radiophonique (chaîne nationale); (iii) discours d'un chef d'état le jour de la fête nationale (télédiffusé ou adressé à un public co-présent); (iv) conférence scientifique (par un orateur devant un public de pairs, lors d'une seule et même conférence); (v) interview radiophonique (émission littéraire); (vi) récit conversationnel. Le *Tableau 1* ci-dessous détaille pour chaque genre de parole et chaque enregistrement: sa durée, son nombre de syllabes et d'unités séparées par des pauses (USP) (voir définition §3.2), et la longueur moyenne en syllabes et en seconde de ces USP.

Selon ce qui a été argumenté sous 2, chaque situation de production est décrite selon une matrice de traits, chaque trait pouvant prendre deux valeurs, ou trois (dans certains cas, une valeur intermédiaire ; dans d'autres, le descripteur est sans objet)[6]. Cette description par trait permet de regrouper certains genres qui partagent les mêmes propriétés situationnelles.

[6] Cette description forme une version simplifiée, et exploitable statistiquement, des 10 critères de Koch & Oesterreicher (2001). Lucci (1983, cité dans Léon 1993:159), retient quant à lui 9 invariants situationnels, globalement opposés les uns aux autres dans les situations « dialogue » (face-à-face), « lecture » (face au micro), et « conférence » (neutralisant les dimensions de sexe, d'âge, de « dialecte »).

Méthodologiquement, cela nous permettra de corréler certains traits prosodiques non pas à un style particulier, mais à une sous-composante d'un genre. La limite de cette description réside d'une part dans le fait que notre corpus ne couvre pas toutes les configurations de traits possibles ; d'autre part, nous ne prétendons pas que chaque échantillon soit « le meilleur représentant » du genre envisagé.

Genre	Enregistrement	Durée (sec)	Nb syll	Nb USP	Nb syllabe par USP	Dur. moyenne USP(sec)
Lecture (LEC)	lec-be	114	606	31	20	3.677
	lec-fr	150	617	42	15	3.571
	lec-ch	137	606	44	14	3.114
Conférence scientifique (CNF)	cnf-ch	219	950	98	10	2.235
	cnf-fr	224	1117	59	19	3.797
	cnf-be	244	1065	46	23	5.304
Interview radio (INT)	int-be	296	1197	91	13	3.253
	int-fr	331	1403	114	12	2.904
Journal parlé (JPA)	jpa-fr	188	971	27	36	6.963
	jpa-be	253	1315	56	23	4.518
	jpa-ch	180	879	42	21	4.286
Récit conversationnel (NAR)	nar-ch	218	948	51	19	4.275
	nar-be	206	945	49	19	4.204
	nar-fr	198	775	44	18	4.500
Discours politique (POL)	pol-be	188	420	66	6	2.848
	pol-ch	230	1011	69	15	3.333
	pol-fr	217	744	84	9	2.583
Total		4183	17799	1013	18	4.129

Tableau 1. Description du corpus, par enregistrement : durée en secondes, nombre de syllabes articulées, et nombres d'unités séparées par des pauses (USP)

Traits situationnels / Genres	Discours médiatique professionnel (M) / non médiatique (NM)[7]	Type d'audience (0 - face à face - beaucoup)	Discours préparé / improvisé
Conférence scientifique	NM	public	semi-préparé (non lu)
Interview radiophonique	M	face à face	semi-préparé (non lu)
Journal parlé	M	0 (micro)	préparé (lu)
Lecture	NM	0 (micro)	préparé (lu)
Récit conversationnel	NM	face à face	improvisé
Discours politique	M[8]	public	préparé (lu)

Tableau 2. Description des genres de parole du corpus selon des traits de la situation de communication

3.2. Présentation des annotations

L'ensemble des échantillons a fait l'objet d'un alignement phonétique avec l'outil de segmentation EasyAlign (Goldman 2010) à partir duquel sont construites différentes couches d'annotations. Le fichier d'annotation complet de chaque échantillon sonore compte 7 *tires* (couches d'annotation) dans le format TextGrid de Praat (Boersma & Weenink 2010) (cf.

[7] « Médiatique » décrit un discours produit par des professionnels des médias, c'est-à-dire un discours qu'on peut qualifier de « journalistique ».
[8] Non médiatique eu égard au (sous-)trait « professionnel » - mais diffusé sur les médias de masse.

Tableau 3). Les 3 premières tires sont produites automatiquement par EasyAlign (phones, syll, words). L'analyse prosodique (décrite ci-dessous) est largement automatisée mais elle recourt à une annotation manuelle de phénomènes liés à la production du discours (prises de souffle, hésitations, interruptions, phénomènes paraverbaux, alternance des tours de parole).

Pour l'analyse prosodique, l'unité de base est la syllabe. A l'aide d'un script basé sur Prosogram, les noyaux syllabiques sont détectés et leur fréquence fondamentale stylisée (Mertens 2004). Cette stylisation vise à minimiser le risque d'erreurs de détection de la f0 pour les mesures mélodiques (Hermes 2006).

Nom de la tire	**Contenu de l'annotation**	
Phones	Transcription phonétique alignée sur le signal, en SAMPA	EasyAlign
Syll	Syllabation de la transcription phonétique	
Words	Découpage en mots graphiques (séparés par une espace ou une apostrophe)	
Delivery	Annotation des phénomènes de production et des proéminences : ! : interruption syntaxique (amorce , faux départ) z : hésitation P : syllabe perçue comme proéminente @ : schwa post-tonique * : prise de souffle audible _ : pause silencieuse % : bruit extérieur + : syllabe avec chevauchement de parole c : creaky voice # : paraverbal	
Gra	Étiquetage (automatique et vérifié manuellement) en catégories grammaticales (VERB, NOUN, ADJ, ADV...)	
Sequences	Découpage manuel en séquences syntaxiques fonctionnelles (Séquence Sujet, Séquence Verbe, Séquence Objet, etc.) (voir Bilger & Campione 2002; Degand & Simon 2009)	
USP	Découpage automatique en unités séparées par des pauses silencieuses longues. Les pauses longues sont isolées.	

Tableau 3. Description des fichiers d'annotation (TextGrids) liés à chaque échantillon du corpus

Chaque syllabe est étiquetée manuellement et selon un protocole strict, comme « proéminente » ou « non proéminente »; une syllabe est proéminente lorsqu'elle se démarque des syllabes environnantes par une durée ou une hauteur moyenne plus importante, ou encore la présence d'un mouvement mélodique interne (Simon et al. 2008). L'annotation en proéminences syllabiques se combine avec une annotation des syllabes ayant une prosodie particulière due au travail de formulation. On identifie ainsi les allongements d'hésitation, les interruptions, etc. De même, les différents types de pauses silencieuses (avec ou sans prise de souffle, etc.) sont spécifiés. Cette annotation prend place dans la tire *delivery*. Elle permet de gérer l'inclusion ou l'exclusion de certains types de syllabes pour la suite de l'analyse.

L'interprétation des syllabes proéminentes dans le système accentuel du français requiert leur localisation par rapport à la chaîne morpho-syntaxique, par rapport à des mots clitiques ou non clitiques et à des constituants grammaticaux (Mertens 1993). Pour cette raison, nous avons ajouté une annotation grammaticale (tire *gra*) automatique et une annotation syntaxique manuelle (découpage en chunks dans la tire *sequences*) (Beaufort 2002, Roekaut 2009).

Chaque enregistrement a été découpé en unités séparées par des pauses (USP) en fonction des *pauses longues* réalisées par le locuteur. On présente souvent le seuil de 180-200 ms comme seuil minimal pour les pauses perçues comme fonctionnelles (Candea 2000: 23-24 ; Lacheret-Dujour & Victorri 2002 définissent l'unité *période* à partir d'une durée de pause de 250 ms). Cependant, au lieu de se fonder sur une durée de pause standard pour opérer un découpage fonctionnel, notre corpus montre qu'il est pertinent d'adapter ce seuil à la parole de chaque locuteur. Les distributions de pauses de la *Figure 1* montrent des structures

bimodales. Cependant, la valeur des pics ou modes (correspondant aux pauses courtes et longues) ainsi que celle des creux, varient selon que l'on considère le corpus entier, seulement un genre (ici POL est représenté) ou chaque locuteur pris isolément. Par exemple, le creux (que l'on retiendra comme valeur-seuil pour discriminer les 2 types de pauses) de pol-be (0.4 s) est bien plus élevé que celui des deux autres locuteurs de ce genre *politique* (230 et 225ms pour respectivement pol-ch et pol-fr).

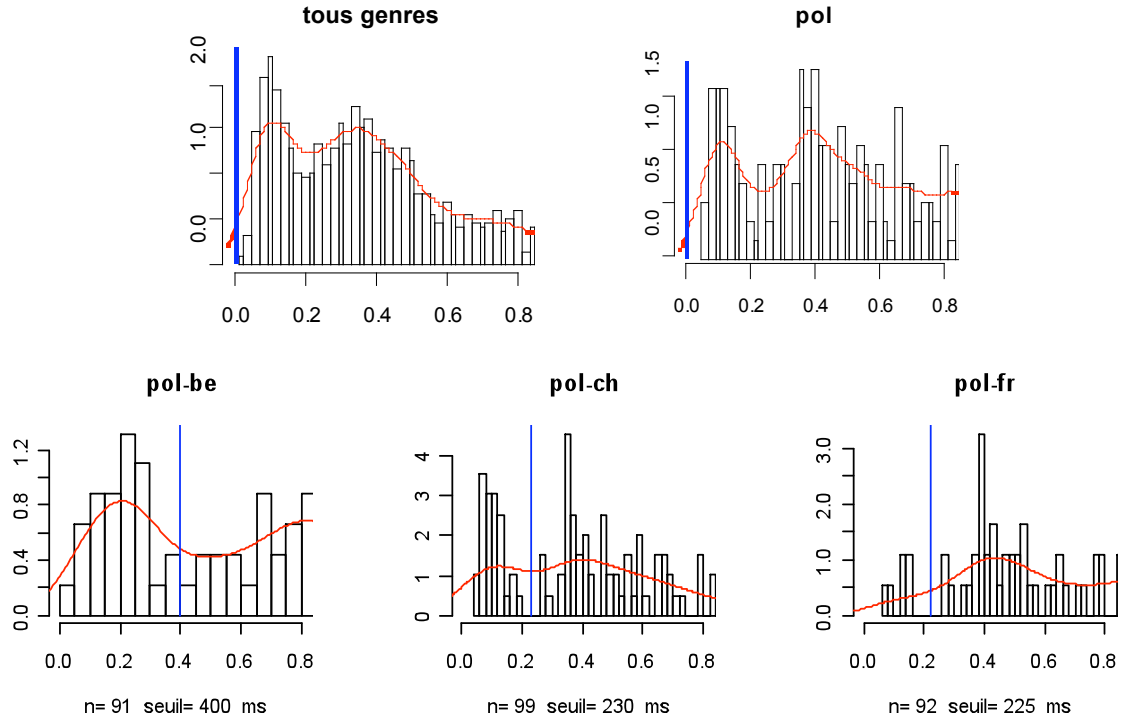

Figure 1. Distributions des durées de pauses (en secondes): tous genres confondus (1732 pauses) ; pour les 3 échantillons de discours politique, puis par échantillon.

Par conséquent, l'analyse de la durée moyenne et de la distribution des pauses[9] pour chaque locuteur a permis d'établir un seuil discriminant des « micro-pauses » et des pauses considérées comme longues (voir *Tableau 4*). Les micro-pauses sont jugées non pertinentes pour le découpage; les pauses longues (supérieures au seuil) permettent de segmenter automatiquement l'enregistrement en USP. Il n'est pas tenu compte du contenu lexico-syntaxique qui précède ou suit ces pauses (par ex. la présence d'un euh d'hésitation, d'une interruption, etc), de sorte qu'on ne différencie pas des pauses « structurantes » ou « d'hésitation » (Candea 2000); la valeur linguistique ou fonctionnelle de ces USP n'est pas analysée ici.

	NAR			POL			INT		LEC			JPA			CNF		
	be	ch	fr	be	ch	fr	be	fr	be	ch	fr	be	ch	fr	be	ch	fr
dur	275	220	275	400	230	225	175	190	260	275	250	200	250	240	270	150	160

Tableau 4. Valeurs seuils (exprimées en ms) pour discriminer les micro-pauses des pauses longues, par locuteur

Le nombre de syllabes des USP et leur durée moyenne peuvent ainsi être calculés (cf *Tableau 1* en §3.1). Ces caractéristiques seront étudiés plus en détail en §4.2.

[9] L'ensemble du corpus totalise 1723 pauses silencieuses.

À partir de ces différents niveaux d'annotation présentés dans la *Tableau 3*, des propriétés acoustiques du signal sont mesurées et mises en regard d'informations linguistiques, comme nous l'expliquons au point suivant.

3.3. Présentation des mesures acoustiques

Les propriétés du signal de parole affectant les syllabes - fréquence fondamentale, durée et intensité - génèrent les caractéristiques prosodiques intonatives, accentuelles et rythmiques observables dans la parole, et perceptibles, voire saillantes. A l'aide de l'outil ProsoReport (Goldman et al. 2007), ces caractéristiques peuvent être mesurées, que ce soit pour l'ensemble d'un enregistrement ou pour des parties de celui-ci.

Les mesures prosodiques obtenues par le ProsoReport et exploitées dans cette étude sont les suivantes:

- **mesures sur les syllabes**: durée moyenne et distribution; hauteur relative; proportion de syllabes proéminentes (selon la détection automatique); ces mesures sont spécifiées pour les différentes catégories de syllabes (syllabes en position initiale ou finale de mot accentuable; autres positions - voir §4.1);
- **mesures sur les unités séparées par des pauses (USP)**: pour chaque USP, on mesure le débit de parole (nombre de syllabes par seconde, y compris les pauses) et d'articulation (en excluant les pauses); l'amplitude du registre tonal (en demi-tons); l'agitation mélodique (en valeur absolue de demi-tons parcourus); la densité accentuelle (proportion de syllabes proéminentes par rapport aux syllabes non proéminentes).

3.4. Hypothèses

D'une part, nous recourons à une approche relativement inductive[10] en cherchant à opérer des regroupements entre les données sonores qui présentent les mêmes caractéristiques prosodiques. Nous nous attendons à ce que les données appartenant aux mêmes genres, ou possédant les mêmes traits situationnels, présentent des caractéristiques prosodiques plus proches entre elles. Cependant, les caractéristiques des locuteurs pourraient résulter dans le fait que deux enregistrements appartenant à deux catégories de genres différentes présentent plus de ressemblance entre eux qu'avec ceux du même genre, pour des raisons par exemple de variation régionale, ou de sexe, ou idiolectale. Plus généralement cela revient à admettre une part de créativité phonostylistique[11].

D'autre part, sur la base d'études préalables (Goldman et al. 2007; Burger & Auchlin 2007; Simon et al. à par.), nous attendons certaines tendances, telles que

- une proportion plus importante d'accents initiaux (voir §4.1) dans le genre journalistique;
- un registre tonal élargi dans les genres de parole publique (discours politique) ou médiatique (journal parlé);
- un débit plus rapide pour la parole lue (lecture; genre journalistique);
- un parcours mélodique global plus important pour le genre médiatique;

[10] Notre approche ne peut cependant pas être qualifiée de *data driven* en raison des annotations basées sur des hypothèses préalables.

[11] Voir par exemple pour l'évolution du phonostyle radiophonique Boula de Mareüil et al. (2008).

4. Résultats

Les résultats sont présentés en deux volets: les mesures sur les syllabes d'abord, les mesures sur les USP ensuite.

4.1. Mesures sur les syllabes

Notre annotation permet, pour chaque échantillon de corpus, de calculer le pourcentage de syllabes proéminentes et de décrire leur localisation (syllabe initiale ou finale de clitique ou de mot accentuable). Dans la mesure où la tire d'annotation *delivery* (voir *Tableau 3*) exclut une série de syllabes (hésitations, faux départ, etc.) et où l'annotation grammaticale indique la catégorie des constituants, on peut postuler que les syllabes proéminentes en position initiale ou finale de polysyllabes des catégories Nom, Adjectif, Verbe[12], Infinitif et Adverbe correspondent respectivement à des accents initiaux et finaux.

Nous nous attendions à voir surgir des différences dans la fréquence des accents initiaux dans les styles médiatiques, par rapport aux autres styles; nous souhaitions explorer si les styles divergeaient quant aux types de constituants affectés d'un accent initial; enfin, nous avons analysé les caractéristiques acoustiques des syllabes selon leur localisation (proéminentes initiales, finales, autres; non proéminentes; voir Astésano 1999: 186-187).

Selon les *Tableau 5* et *Tableau 6*, les trois styles médiatiques se distinguent globalement par un nombre plus important de syllabes proéminentes (33% en moyenne contre 29% pour les styles non médiatiques). Cette différence est accrue si on considère le pourcentage de syllabes proéminentes en position initiale de mot accentuable polysyllabique : dans les styles médiatiques, 28% des mots pleins portent un accent initial, contre 15% pour les styles non médiatiques (alors qu'ils ont le même nombre de syllabes). Le degré de préparation d'un discours semble être un autre trait qui favorise l'apparition de ce type d'accent: 15% de mots pleins portent un accent initial dans les styles non préparés (conversationnel), 18% dans les styles semi-préparés et 24% dans les styles préparés et lus. Le trait situationnel qui décrit le type d'audience ne semble pas avoir d'impact sur cette variable. Globalement, la variation va dans le même sens pour les accents finals (voir dernière colonne du *Tableau 6*).

Genre	Prom (%)	I	F
nar	25	14	48
lec	26	13	54
cnf	31	15	**64**
pol	**31**	**25**	**61**
jpa	**33**	**31**	51
int	**34**	**28**	58

Tableau 5. Proportion de syllabes proéminentes par genre, et selon leur localisation grammaticale (I-initiale ou F-finale de mot accentuable)

[12] à l'exception des auxiliaires.

Trait situationnel		Prom (%)	I	F
Audience	micro	30	22	53
	face à face	30	19	51
	public	31	19	**63**
Médiatique	non-médiatique	29	15	56
	médiatique	33	28	56
Préparé	non préparé	28	**15**	51
	semi-prép	32	**18**	63
	préparé et lu	31	**24**	55

Tableau 6. Proportion de syllabes proéminentes par variable situationnelle, et selon leur localisation grammaticale (I-initiale ou F-finale de mot accentuable)

Le *Tableau 7* détaille ces fréquences selon la catégorie grammaticale du mot affecté d'une proéminence en syllabe initiale. Il apparaît que les genres « lecture » et « narration conversationnelle » sont relativement proches en ce qu'ils contiennent le moins de syllabes initiales accentuées. Le genre « journal parlé » se caractérise par une distribution égale des accents initiaux entre les quatre catégories grammaticales, tandis que ces accents affectent essentiellement les adverbes et les verbes dans le genre « conférence » et les adjectifs dans le genre « politique ».

		ADV	ADJ	VERB	NOM
Genre	cnf	32	9	26	13
	jpa	22	25	33	32
	nar	12	21	12	12
	int	31	50	24	29
	pol	43	38	31	19
	lec	11	16	12	12

Tableau 7. Pour chaque catégorie grammaticale, et uniquement pour les polysyllabes: pourcentage de mots affectés d'une proéminence initiale.

La distribution des accents initiaux selon les traits situationnels résiste quelque peu à notre analyse. Le trait du type d'audience (seul face à un micro, face à un public…) ne semble pas corrélé à la distribution des accents initiaux. Les discours médiatiques présentent en moyenne deux fois plus d'accents initiaux que les discours non médiatiques, sans que cette augmentation soit spécifique pour une des catégories grammaticales. Enfin, les discours non préparés se distinguent à la fois des discours semi-préparés et préparés par un nombre significativement plus faible d'accents initiaux.

		ADV	ADJ	VERB	NOM
Audience	micro	16	22	18	23
	face à face	17	31	15	19
	public	36	20	29	15
Médiatique	non-médiatique	18	14	15	13
	médiatique	32	31	30	28
Préparé	non préparé	12	22	14	14
	semi-prép	32	14	25	17
	préparé et lu	28	26	27	23

Tableau 8. Pourcentage de mots affectés d'une proéminence initiale par traits situationnels

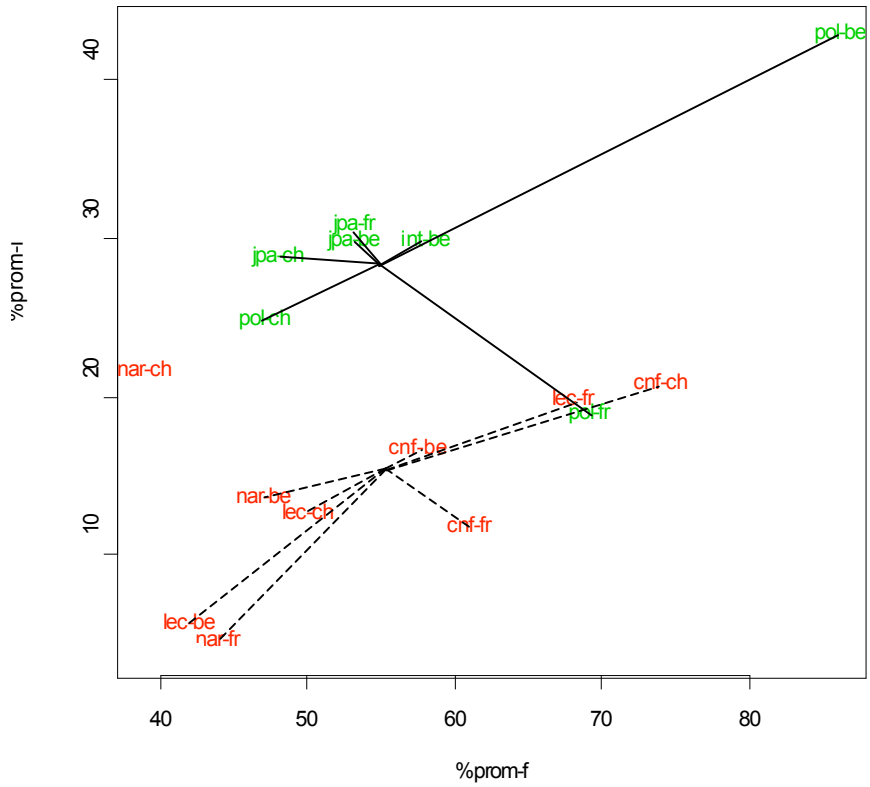

Figure 2. Distribution des locuteurs selon la proportion de proéminences finales et initiales. En rouge les genres non médiatiques, en vert les genres médiatiques

4.2. Mesures sur les unités séparées par des pauses (USP)

Le séquençage de chaque enregistrement en USP vise, entre autres, à mesurer l'homogénéité des caractéristiques prosodiques sur sa durée, c'est-à-dire la constance du style. Précédemment (Goldman & al. 2007; Goldman & al. 2008; Simon et al. 2010), nous avons décrit chaque phonostyle à l'aide d'un rapport prosodique global, c'est-à-dire de mesures moyennes sur un enregistrement (débit moyen, registre moyen, etc.). La question se pose pourtant de savoir si un locuteur est constant dans sa manière d'exploiter les traits prosodiques.

Que ressort-il de l'observation des propriétés prosodiques des 1013 USP du corpus, quand on compare les locuteurs et les paramètres situationnels ? Cette analyse vise à vérifier si les (familles de) styles se discriminent entre eux sur la base de caractéristiques prosodiques, et aussi à analyser les éventuelles ressemblances inter-genres (voir la description par paramètres). Selon un troisième cas de figure, nous envisageons que les prosodies de certains locuteurs se ressemblent, malgré qu'ils ne partagent pas la même situation de communication.

La première observation concerne la longueur moyenne des USP (en nombre de syllabes) dans chaque genre. Le style JPA présente les USP les plus longues (en moyenne 22 syll/USP) tandis que les styles INT et POL ont les USP les plus brèves (autour de 10 syll/USP). Entre ces valeurs extrêmes, on trouve les autres styles NAR, CNF et LEC.

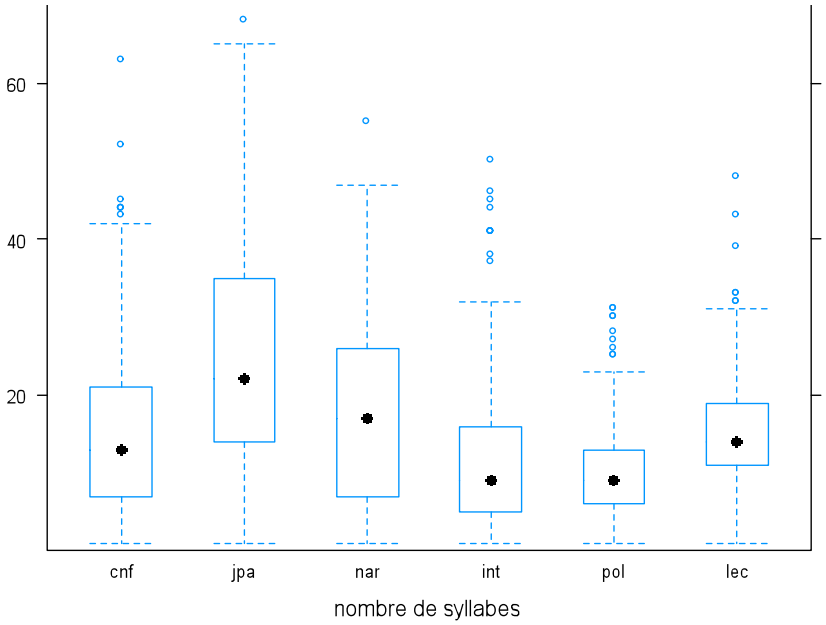

Figure 3. USP: nb de syll / genre

Aucun des traits situationnels (type d'audience, degré de préparation ou médiatique vs. non médiatique) n'est corrélé à la longueur des USP. Le fait que le style « journal parlé » ait les USP les plus longues peut s'expliquer par le caractère très rapide du débit et la présence essentiellement de pauses prises de souffle. La brièveté des USP dans le style POL peut s'expliquer par le caractère solennel de ce type de discours, tandis qu'elle s'explique plutôt par la planification nécessaire à la production du discours dans les interviews radiophoniques (voir Degand & Simon 2009 qui analysent le placement des pauses et des frontières prosodiques par rapport aux structures syntaxiques). Aucun trait situationnel n'est commun aux styles se caractérisant par des USP de longueur moyenne (CNF, NAR et LEC).

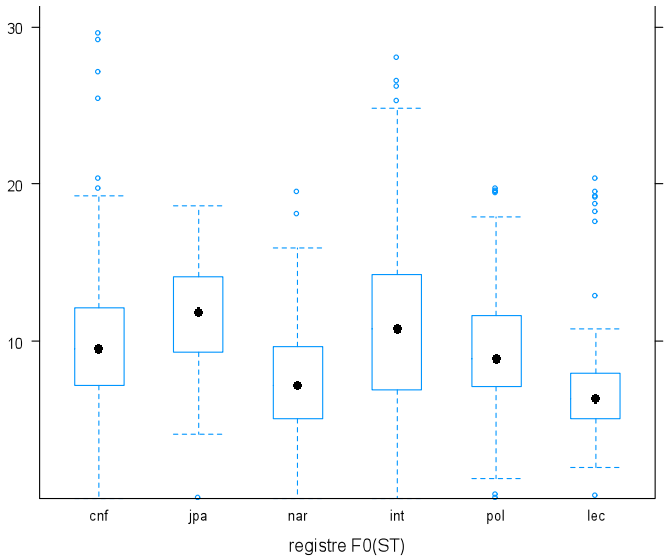

Figure 4. USP : amplitude du registre mélodique (en demi-tons), par genre

La *Figure 4* décrit l'amplitude du registre mélodique exploité par le locuteur. Il apparaît très clairement que les styles publics présentent des registres de F0 plus amples / élargis; cet effet est particulièrement accru pour les styles radiophoniques (environ 13 demi-tons pour

JPA et INT et 10 demi-tons pour POL et CNF). L'étendue du registre est réduite pour LEC et, dans une moindre mesure, pour NAR. Nos données confirment donc les observations de Blaauw sur le néerlandais (cité par Llisterri 1992: 14) selon lesquelles le registre de f0 est réduit en lecture par rapport à la conversation spontanée.

Enfin, les mesures de débit de parole (avec les pauses silencieuses) et d'articulation (sans les pauses silencieuses) distinguent les styles lus des styles non préparés, à l'exception des discours politiques.

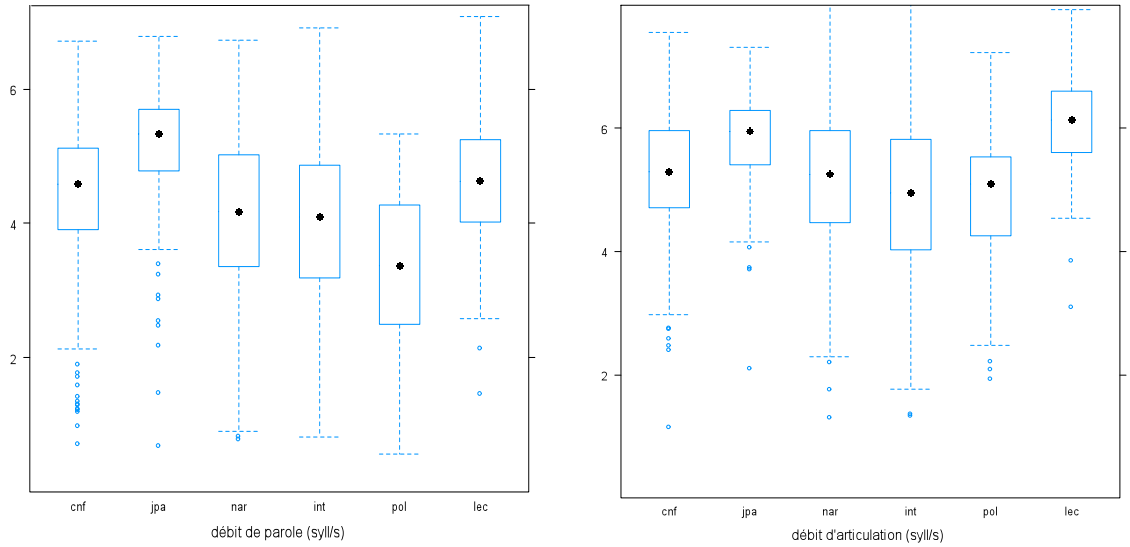

Figure 5. USP : débit de parole et d'articulation (sans les pauses silencieuses, en syll.sec), par genre (en syll/sec)

Pour ce qui est du débit d'articulation, il est plus élevé pour les styles lus LEC et JPA et pour ces deux styles la distribution du débit est très stable, mais pas pour POL qui présente un débit particulièrement lent : on peut attribuer ce trait au caractère solennel de ce type de discours – mais il faut noter que les 3 échantillons de parole ne se comportent pas de manière tout à fait homogène. Le trait situationnel que nous avons retenu ne rend donc pas compte de ces différences, qui peuvent être imputées également à des contraintes liées à la rentabilité des médias (pour la vitesse d'élocution des journalistes.

Pour ce qui est du débit de parole (incluant les pauses longues), JPA se différencie encore plus nettement (à cause du petit nombre et de la faible durée des pauses) d'un côté et POL de l'autre (pauses longues et nombreuses). Les autres styles ont des valeurs proches.

5. Conclusion

5.1. Discussion des résultats

Ce travail décrit la mise en place d'une méthodologie d'analyse prosodique outillée dans le but de comparer les réalisations prosodiques de différents locuteurs, dans des genres contrastés distingués les uns des autres en termes de traits situationnels. Deux objets principaux sont examinés : i. la proéminence syllabique, son taux, et sa localisation dans le mot, et par catégorie grammaticale ; ii. les unités séparées par des pauses, déterminées par un seuil de pause ajusté au locuteur, décrites par des mesures classiques (nombre de syllabes par USP, registre, débit).

Nos observations confirment l'hypothèse d'une proportion importante de proéminences initiales pour le trait situationnel « médiatique » (15% vs 28%, représenté graphiquement figure 2), alors que le trait « audience » ne semble pas influencer ce taux (*Tableau 6*).

Les observations par USP permettent une description plus riche qu'une moyenne globale par enregistrement, donnant accès à la distribution, régulière ou irrégulière, des valeurs dans le temps (constance et régularité des locuteurs). Ainsi la lecture à haute voix, réputée pour avoir un registre tonal réduit, s'avère également constante dans cette dimension, à l'opposé de l'interview radio qui présente un registre tonal plus ample, mais plus irrégulier. De même, le débit moyen élevé du journal parlé varie peu en comparaison avec le discours politique, ce qui peut être interprété par un « effet plafond ».

Deux écueils se présentent sur la voie de la définition des phonogenres, le fait que certains sont compacts, et d'autres beaucoup moins, et le fait qu'il n'est pas vraiment possible d'énumérer un jeu de traits définitoires. Par ailleurs, pour distinguer plus formellement les influences respectives du phonogenre et de l'idiostyle qui interviennent dans la production orale, il convient de prendre en considération un nombre plus élevé d'échantillons pour chaque genre envisagé, ainsi que de multiplier les genres pris en considération, et donc les traits situationnels.

5.2. La variation prosodique - phonostylistique, quel intérêt?

Ce questionnement présente un intérêt strictement prosodique: disposer de mesures de variations pour une population donnée; et un intérêt sociolinguistique discursif: disposer de mesures (certes, peu documentées) croisées sur différentes conditions de parole et différentes régions linguistiques francophones (échantillonnage non représentatif de la francophonie).

Le dernier intérêt est d'ordre pragmatique et épistémologique, et réside dans le sens que l'on donne à cette variation. Les différences constatées ne servent pas à communiquer de la signification, des concepts; elles servent, globalement, la « fonction identificatrice »: que le parler reflète un locuteur singulier, un rôle typifié, une situation ou un ingrédient spécifique de celle-ci. La « fonction identificatrice » agit dans deux directions:

- pour le producteur de la parole, elle projette et lui permet de contrôler son identité de parole, par conformité à un genre normé ou singularisation;
- pour le récepteur, elle consiste à l'informer quant à la source de la parole. Cette information elle-même établit un chemin direct entre des perceptions auditives et la perception-identification, pré-conceptuelle, d'une identité, individuelle ou générique. L'expérience ordinaire permet de supposer que cette perception catégorise prématurément le familier / non-familier, le standard / singulier.

Ainsi, la variation prosodique détermine des dimensions et des qualités de l'expérience de parole, du locuteur, et de l'auditeur.

Références

Astésano, C. (1999). *Rythme et discours: invariance et sources de variabilité des phénomènes accentuels en français*. Thèse de doctorat de Sciences du Langage, Aix-en-Provence: Université Aix-Marseille I.

Beaufort, R., T. Dutoit & V. Pagel (2002). Analyse syntaxique du français. Pondération par trigrammes lissés et classes d'ambiguïtés lexicales. In *Proceedings of JEP*, pp. 133–136.

Bilger, M. & E. Campione (2002). Propositions pour un étiquetage en 'séquences fonctionnelles'. *Recherches sur le français parlé* 17, pp. 117-136

Boersma, P. & D. Weenink (2010). Praat: doing phonetics by computer (Version 5.1.29) [Computer program]. Retrieved March 11, 2010, from http://www.praat.org/

Boula de Mareüil, A. Rilliard & A. Allauzen (2008). A diachronic study of prosody through French audio archives. *4th Conference on Speech Prosody*, Campinas, pp. 531–534.

Burger, M. & A. Auchlin (2007). Quand le parler radio dérange : remarques sur le phono-style de France Info. Broth, M., M. Forsgren, C. Norén & F.Sullet-Nylander (éds). *Le Français parlé des médias. Actes du colloque de Stockholm 8-12 juin 2005*, Acta Universitatis Stockholmiensis, Stockholm, pp. 97-111.

Callamand, M. (1987). Aspects prosodiques de la communication. *Études de linguistique appliquée* 66, Paris, Didier.

Campbell, N. (2004). Accounting for voice-quality variation. Speech Prosody 2004, 217-220

Candea, M. (2000). *Contribution à l'étude des pauses silencieuses et des phénomènes "d'hésitation" en français oral spontané. Etude sur un corpus de récits en classe de français*. Thèse de Doctorat, Université Paris III.

Degand E. & A. C. Simon (2009) On identifying basic discourse units in speech: theoretical and empirical issues. *Discours 4* [En ligne]. URL : http://discours.revues.org/index5852.html.

Eskenazi, M., (1993). Trends in Speaking Styles Research. *ISCA*, pp. 501-509.

Fónagy, I. & J. Fónagy (1976). Prosodie professionnelle et changements prosodiques. *Le Français Moderne* 44, pp.193-228.

Fónagy, I. (1983). *La vive voix. Parole et expressivité*. Payot, Paris.

Goldman J.-P. (2010) EasyAlign [Computer program]. Retrieved June 18, 2010 from http://latlcui.unige.ch/phonetique.

Goldman J.-P. & al. (2007). Phonostylographe : un outil de description prosodique. Comparaison du style radiophonique et lu. *Nouveaux cahiers de linguistique française* 28, pp. 219-237

Goldman J.-P. & al. (2008). ProsoReport: an automatic tool for prosodic description. Application to a radio style. Speech Prosody -2008, pp. 701-704.

Goldman, J.P., T. François, S. Roekhaut & A. C. Simon (2010). Étude statistique de la durée pausale dans différents styles de parole, *Journées d'Etudes sur la Parole*, Mons, Belgique

Hermes, D.J. (2006). Stylization of Pitch Contours. Sudhoff S. & al. (eds), *Methods in Empirical Prosody Research*. Berlin and New York, Walter De Gruyter, pp. 29-61.

Hirschberg, J., 2000. A corpus-based approach to the study of speaking styles. Horne, M. (ed). *Prosody, Theory and Experiment: Studies Presented to Gösta Bruce*. Amsterdam, pp. 335-350.

Johns-Lewis C. (1986). The prosodic differentiation of discourse modes. Johns-Lewis C. (ed.), *Intonation in Discourse*, Croom Helm, London & Sidney, pp. 199-219.

Koch, P. & W. Oesterreicher (2001). Langage parlé et langage écrit. Holtus, G., M. Metzeltin, Ch. Schmitt (eds). *Lexicon der Romanistischen Linguistik*, tome 1-2, Max Niemeyer, Tübingen, pp. 584-627.

Koopmans-van Beinum, F. (1991). Spectro-temporal reduction and expansion in spontaneous speech and read text : Focus words versus non-focus words. In *Proceedings of the ESCA workshop Phonetics and Phonology of Speaking Styles: Reduction and Elaboration in Speech Communication*, Barcelona, paper 036.

Lacheret-Dujour, A. & B. Victorri (2002). La période intonative comme unité d'analyse du français parlé: modélisation prosodique et enjeux linguistiques. *Verbum* XXIV/1-2, pp. 55-72.

Léon, P. (1993). *Précis de phonostylistique. Parole et expressivité*. Nathan Université, Paris.

Llisteri, J. (1992). Speaking styles in speech research. *ELSNET/ESCA/SALT Workshop on Integrating Speech and Natural Language*, Dublin, Ireland. http://liceu.uab.es/~joaquim/publicacions/SpeakingStyles_92.pdf [consulté 20 juin 2010].

Lucci, V. (1983). *Etude phonétique du français contemporain à travers la variation situationnelle*. Thèse de Doctorat, Publications de l'Université de Grenoble.

Mertens, P. (1993). Accentuation, intonation et morphosyntaxe. *Travaux de Linguistique* 26, pp. 21-69.

Mertens, P. (2004). Le Prosogramme: une transcription semi-automatique de la prosodie. *CILL* 30, no 1-3, 7-25.

Roekhaut, S. (2009). *Expressive. Système automatique de diffusion vocale d'information dédicacée: synthèse de la parole expressive à partir de textes balisés*, Scientific Report (Convention n° 0616422 avec la Région wallonne), Unpublished ms.

Simon A. C., M. Avanzi, J.-P. Goldman (2008). La détection des proéminences syllabiques. Un aller-retour entre l'annotation manuelle et le traitement automatique. *Congrès Mondial de Linguistique Française 2008*, Paris, Juillet 2008

Simon, A.C. et al. (2010). Les phonostyles: une description prosodique des styles de parole en français. in Abécassis M. & G. Ledegen (eds), *Les voix des Français : en parlant, en écrivant*, Bern, Lang, pp. 71-88.