Indices prosodiques caractérisant un style d'élocution et ses variations individuelles

Jean-Sylvain Liénard - Martine Adda-Decker

LIMSI-CNRS
Bât. 508, BP 133, 91403 Orsay cedex, France
Tél. : +33 (0)1 69 85 81 13 - Fax : +33 (0)1 69 85 80 88
Mél : jean-sylvain.lienard@limsi.fr - http://www.limsi.fr/MIDL/

ABSTRACT

From the point of view of prosody, Automatic Identification of Language, as well as dialectal studies, have to deal with the variability problem. Thus we propose to investigate which aspects of the signal are related to a given speech style, and which ones are related to the particular voice and speaking manner of a given speaker. In our corpus a French text is read aloud twice by 48 female speakers. We consider as a primary prosodic unit the Vocalic Group, extracted from the signal without using any knowledge of its linguistic content. Then we study the pause distribution and we align all the utterances on the same time scale. This processing yields an average prosodic representation, which can be attributed to the pronunciation norm observed by all of the speakers. The individual variations are characterized by global shifts in the mean value of each prosodic feature, as well as by their evolution along the utterance. Eventually the proposed approach permits to determine in what respect, where and how much a given utterance deviates from the norm.

1. INTRODUCTION

La multiplicité des sources d'information présentes dans tout signal de parole (contenu linguistique, mais aussi style de parole, force de voix, spécificité du locuteur, conditions de la communication) fait qu'il est difficile d'évaluer objectivement la contribution de l'une ou de l'autre, en particulier si l'on considère l'ensemble des paramètres (segmentaux et prosodiques) du signal [1]. L'identification automatique des langues, accents et dialectes n'échappe pas à ce problème de variabilité. C'est pourquoi il nous semble nécessaire, avant même de tenter d'identifier une langue ou une variante régionale, de préciser autant que possible dans quels paramètres du signal on peut trouver les informations propres au message, au style d'élocution [2], au groupe de locuteurs considéré, ou à chaque locuteur pris individuellement.

Dans la présente étude nous nous limitons à la lecture orale d'un petit texte français par un ensemble homogène de locutrices. Nous cherchons à mettre en évidence l'existence d'une structure prosodique commune reflétant le style de parole adopté par les locutrices, et à définir ses caractéristiques acoustiques. Puis nous cherchons dans quelle mesure chacune s'écarte du modèle commun, et dans quels aspects du signal se manifestent ces écarts. L'objectif n'est pas tant de découvrir de nouvelles propriétés de ce style d'élocution, bien connu et exploré en phonétique, que de définir une méthode générale permettant d'étudier les invariants prosodiques et les variations attachés à un même contenu linguistique, sans connaître le détail de ce contenu.

2. CORPUS ET TRAITEMENTS

2.1 Le corpus

Le corpus mis à notre disposition par M.A. Leblanc [3] comprend la lecture orale d'un petit texte de Pierre Daninos (tableau 1) par 48 étudiantes de psychologie de l'Université Paris X Nanterre, francophones, âgées de 20 à 25 ans. Le texte comprend 5 phrases, présentées en un seul paragraphe. Il est lu deux fois par chaque locutrice, sans autre consigne que de faire attention au sens du texte. La durée de chaque séquence varie selon la locutrice entre 20,5 et 34,1 secondes. Les conditions d'enregistrement (en studio) sont sensiblement les mêmes pour toutes. On n'a pas cherché à obtenir une élocution parfaite: certaines séquences comportent quelques hésitations, reprises, bafouillages. Ce corpus de 96 séquences constitue une instanciation du style "lecture à voix haute" tel que nous l'avons tous appris et pratiqué sur les bancs de l'école primaire. Il est immédiatement repérable à l'écoute par tout auditeur francophone.

TAB. 1 – *Texte lu par les locutrices du corpus Leblanc.*

Les Français, qui consacrent une partie appréciable de leur journée à la poignée de main, passent également un temps considérable à se prier réciproquement d'entrer dans leurs maisons. Les uns prient les autres d'entrer, les autres jurent qu'ils n'en feront rien. Les premiers disent : " Moi non plus ". Et, de fil en aiguille, les Français ont passé (environ) trois siècles et demi depuis Charlemagne sur le pas de leurs portes. On est même étonné d'en trouver quelques-uns chez eux.

2.2 Détection des noyaux vocaliques

Pour extraire une structure prosodique commune nous avons d'abord tenté d'étiqueter automatiquement chaque séquence. Diverses expériences d'alignement automatique de chaque séquence sur le texte phonétisé nous ont montré la difficulté de la tâche. Le problème réside dans le fait que la transcription phonétique du

texte est une opération normative, alors que la réalisation du texte est spécifique à chaque locutrice et diffère légèrement d'une séquence à l'autre. Tout forçage est artificiel, que la transcription soit automatique ou manuelle, que les entités considérées soient des phonèmes, des syllabes ou des mots, parce que les niveaux d'abstraction considérés sont différents. Ceci nous a incités à adopter une démarche délibérément ascendante: au lieu de chercher en premier lieu à plaquer l'information linguistique sur le signal, nous partons du signal sans utiliser l'information linguistique associée.

S'agissant de prosodie, les parties du signal qu'il convient de détecter sont essentiellement les voyelles. Plus précisément nous nous intéressons aux Noyaux Vocaliques (NV), ceux-ci pouvant comprendre une ou plusieurs voyelles ou diphtongues, voire même plusieurs syllabes du texte réduites en un seul segment de nature vocalique par le fait d'une prononciation rapide ou négligée.

La détection des noyaux vocaliques peut être faite de manière approximative à partir de l'énergie du signal dans la bande basse du spectre (au dessous de 1 kHz), de façon à ne pas prendre en compte l'énergie des consonnes fricatives et plosives dans leur phase de constriction. Dans la suite nous appellerons Intensité du noyau vocalique la mesure en dB relevée à l'instant du maximum de cette fonction énergie BF, définie avec une résolution temporelle de l'ordre de 50 ms. Cette opération, comme tous les traitements suivants, est effectuée au moyen du logiciel PRAAT [4]; pour l'intensité la résolution temporelle est réglée implicitement par le paramètre "f0min" de l'analyse, fixé ici à 40 Hz.

2.3 Détection des pauses

Les phrases du texte écrit sont séparées par des pauses silencieuses par toutes les locutrices, mais il existe de grandes variations en ce qui concerne les autres ponctuations, traduites par des pauses ou par des modifications prosodiques locales, ou simplement ignorées.

Nous considérons comme pause toute chute d'intensité d'au moins 12 dB par rapport à la moyenne de la séquence et durant plus de 1,8 fois la durée moyenne séparant deux noyaux vocaliques, celle-ci étant calculée en ne tenant pas compte des interruptions de plus de 300 ms. Ces valeurs ne sont pas critiques, et dans la majorité des cas les pauses se distinguent nettement des autres segments.

2.4 Indices prosodiques locaux

Les indices retenus sont au nombre de quatre. L'intensité (int) a été mentionnée plus haut. En ce qui concerne F0, celle-ci est calculée avec une moindre résolution temporelle (de l'ordre de 100 ms, soit "f0min" fixé à 15 Hz). Ceci vise à minimiser les erreurs de mesure, inévitables malgré la qualité de l'algorithme de PRAAT, qui effectue un post-traitement par programmation dynamique. Notre stratégie, qui consiste à retenir pour F0 la valeur mesurée à l'instant d'intensité maximale du noyau vocalique, contribue elle aussi à réduire les erreurs, car celles-ci sont le plus souvent observées lors des parties faibles ou transitoires du signal.

F0 est exprimée en demi-tons (st) par rapport à 100 Hz. La raison en est que nous visons à comparer des valeurs produites par différents locuteurs, et que cette opération nous semble plus légitime si elle est effectuée selon une échelle logarithmique, plutôt que selon l'échelle linéaire (en Hz) habituelle.

Retenir une seule valeur de F0 par NV peut paraître insuffisant pour caractériser les variations locales. Nous définissons un second indice (dF0) égal à la différence de F0 observée entre le début et la fin du NV. Les instants de début et de fin sont définis à -3 dB par rapport à l'instant d"intensité maximale.

Enfin nous associons à chaque NV une mesure de durée (dur). Pour cela nous définissons une autre entité segmentale nommée Groupe Vocalique (GV), qui contient le noyau vocalique. Pour éviter les incertitudes de mesure que l'on peut rencontrer dans les faibles niveaux des intervalles intervocaliques, les frontières de GV sont définies comme le milieu des intervalles intervocaliques précédant et suivant le noyau vocalique considéré, en excluant les pauses. Il est à noter que les GV ne sont pas des syllabes, pas même des "syllabes acoustiques", puisque toute référence à une structuration en consonnes et voyelles est ignorée.

Chaque séquence est donc ramenée à une succession de segments, qui représentent des groupes vocaliques et des pauses, élaborés sur la base de traitements purement acoustiques. Chaque segment est caractérisé sous l'aspect prosodique par un ensemble de valeurs prises par les quatre indices définis ci-dessus.

2.5 Alignement sur une séquence de référence

Bien que le texte lu soit le même pour toutes les locutrices le nombre de segments résultant de l'analyse acoustique est variable, selon de débit et le soin apporté à l'articulation. Pour comparer et moyenner les valeurs des indices prosodiques segmentaux il faut être sûr que l'on compare des segments correspondant au même rang séquentiel dans le texte, et nous voulons effectuer cette opération sans être obligés de fournir au système la transcription phonétique réaliste du signal. Une solution est l'alignement de chaque séquence sur une même séquence de référence.

Dans cette étude le choix de la séquence de référence a été effectué à la suite d'une écoute critique de toutes les séquences. Nous avons écarté les séquences trop rapides, trop lentes, hachées par des pauses trop

nombreuses, altérées par des hésitations, des reprises ou des erreurs de lecture, articulées de manière relâchée, marquées d'un accent régional, manifestant une expressivité excessive ou une mauvaise compréhension du texte. Nous avons choisi arbitrairement comme séquence de référence l'une des séquences restantes, au demeurant encore assez nombreuses (environ la moitié du corpus). Dans les exemples qui suivent la séquence de référence est celle qui porte le numéro 07.

L'alignement a été effectué par programmation dynamique sur des critères spectraux: les 8 premiers mfcc, calculés pour chaque instant central d'un noyau vocalique, sur une fenêtre temporelle large (30 ms).

2.6 Représentations prosodiques alignées

La figure 1 illustre le résultat obtenu en moyennant l'ensemble des 96 séquences du corpus, alignées sur la séquence de référence. De haut en bas sont représentés les graphes des quatre indices prosodiques F0, dF0, Intensité et Durée, du début à la fin de la séquence. Les valeurs prises par les indices pendant les pauses ne sont pas représentées, de façon à séparer visuellement les constituants de la séquence.

Nous avons reporté en abscisse une transcription orthographique approximative (avec les élisions, insertions et liaisons) associée à chaque groupe vocalique de la séquence de référence. On notera que cette transcription n'est pas parfaite au sens normatif du texte écrit: on voit apparaître tant des éclatements d'un GV en plusieurs segments, que des segments réduits correspondant à plusieurs syllabes, ou des segments qui ne comportent pas de noyau vocalique.

3. STRUCTURE PROSODIQUE MOYENNE

La représentation segmentale alignée décrite ci-dessus permet de faire quelques remarques sur la structure prosodique moyenne des séquences, que l'on peut considérer comme caractéristique du style de lecture orale mis en œuvre par les locutrices.

3.1 Pauses

Les pauses majeures réalisées dans la séquence de référence sont observées par la très grande majorité des locutrices. Les informations concernant les pauses supplémentaires, les dédoublements et les réductions des groupes vocaliques sont disponibles dans les représentations segmentales individuelles avant alignement. On peut ainsi déterminer en certains points du texte une probabilité de pause qui correspond sensiblement à la présence des ponctuations mineures et des parenthèses du texte.

Le rapport entre le temps total de pause et le temps total de parole traduit un certain équilibre entre la rapidité de l'élocution et le souci de laisser à l'auditeur le temps nécessaire pour comprendre ce qui est prononcé. Ce rapport s'établit en moyenne à 15,6 %; il varie assez largement selon les locutrices (entre 7,1% et 25,7%); il est peu dépendant de la durée totale de la séquence et du débit moyen.

3.2 Double ligne de déclinaison de F0

F0 décroît globalement du début à la fin du texte, selon un phénomène mis en évidence par J. Pierrehumbert en 1981 [6], et observé depuis dans de nombreuses langues. Ce phénomène est apparent sur la séquence moyenne; une analyse de régression linéaire fait apparaître une baisse moyenne de 1,5 demi-ton entre le début et la fin du texte, soit une pente de l'ordre de 0,07 demi-ton par seconde.

La figure 1 montre également une autre forme de déclinaison de F0, par grand groupe interpausal, avec une pente beaucoup plus sensible, pouvant atteindre 1 demi-ton par seconde. Ce phénomène a été observé notamment par N. Grönnum [7] et se trouve largement confirmé dans notre corpus.

3.3 Groupements hiérarchiques

Tous les indices, et en particulier F0, font apparaître des groupements prosodiques de diverses tailles. A l'exception de deux d'entre eux, qui sont plus courts et s'apparentent à des incises, chacun des 7 groupes interpausaux se caractérise par une décroissance de F0 en moyenne, une évolution notable de F0 sur les deux derniers GV (en général une forte décroissance), une décroissance de l'intensité en moyenne, et une décroissance moins marquée de la durée des GV. Tout se passe comme si chaque groupe interpausal, en moyenne décroissant selon tous les indices, était muni d'un marqueur final, qui à la fois annonce et précise la signification de la pause qui suit.

A l'intérieur du groupe interpausal, on observe une succession de groupes prosodiques de petite taille (entre 2 et 4 GV), qui se caractérisent essentiellement par une montée de F0 suivie d'une descente, et secondairement par une augmentation locale de la durée. Ces groupes peuvent être qualifiés de "mots prosodiques" et sont structurés par un ou plusieurs accents manifestés par F0 et Durée, selon une combinaison qu'il convient d'approfondir et qui pourrait caractériser une langue ou un dialecte.

Entre les deux on observe des groupements moins bien définis, qui n'ont pas, dans notre corpus, le caractère systématique des deux types précédents, si ce n'est qu'ils apparaissent comme des subdivisions des groupes interpausaux longs. Ces groupements sont en rapport avec le contenu linguistique du signal et font l'objet principal des études contemporaines sur l'intonation (voir par exemple[7]).

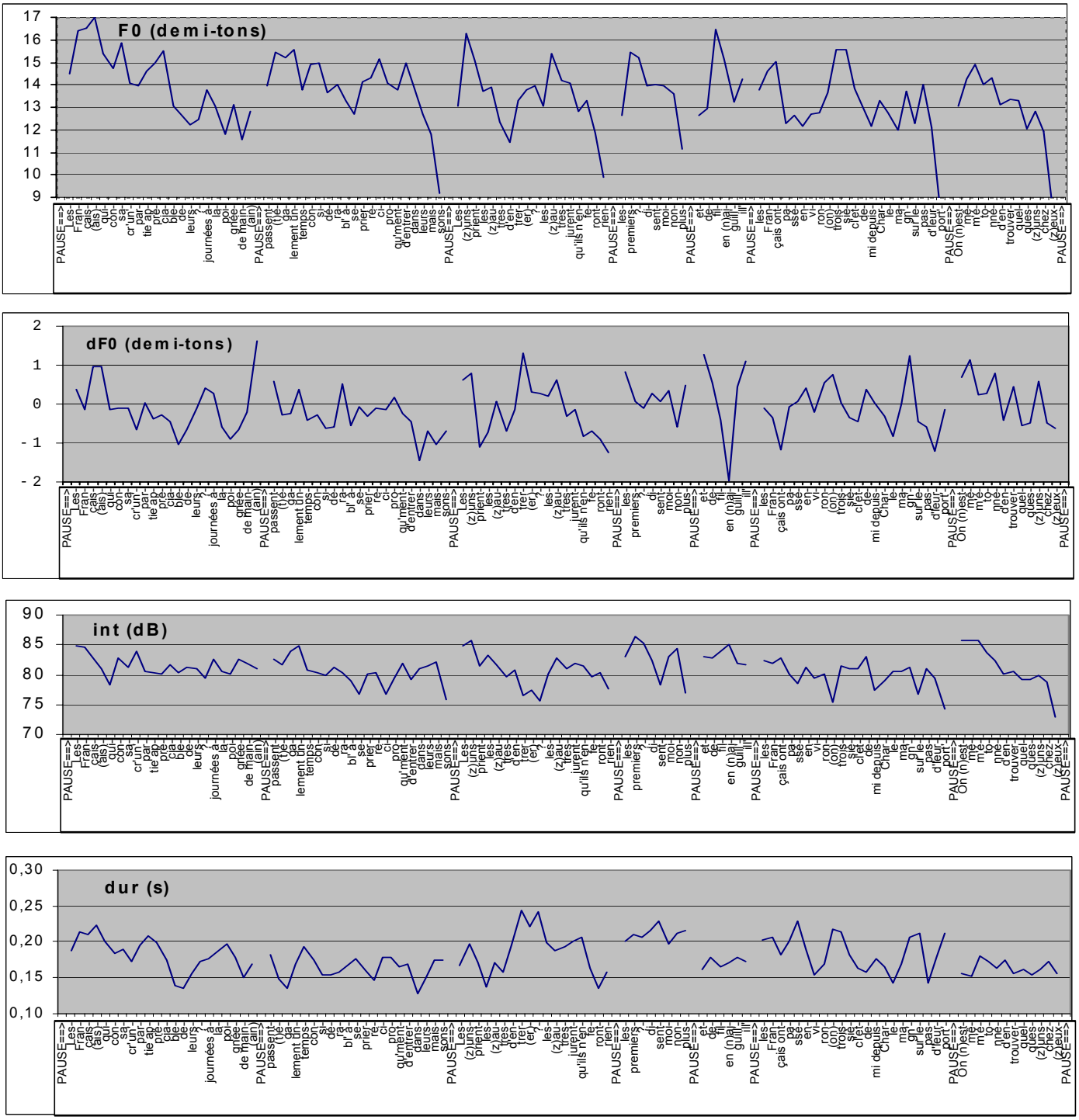

FIG. 1 – *Evolution des indices prosodiques sur la moyenne des 96 séquences alignées. De haut en bas: F0 (demi-tons), dF0 (demi-tons), Intensité int (dB), Durée des groupes vocaliques dur (s)*

.

4. VARIATIONS INDIVIDUELLES

Pour examiner les variations prosodiques individuelles il suffit de comparer les tableaux d'indices soit entre eux, soit par rapport à la moyenne. Aux indices prosodiques nous ajouterons, pour comparaison, les indices spectraux que constituent les quatre premiers des 8 mfcc utilisés lors de l'alignement. L'objectif est d'évaluer la capacité de chaque indice à représenter la spécificité, prosodique ou spectrale, de chaque séquence.

4.1 Mesures de dissemblance

Pour chaque indice nous distinguerons deux mesures de dissemblance entre séquences. La première est la différence entre les valeurs moyennes prises par l'indice considéré pour chacune des deux séquences. Pour clarifier les termes, nous parlerons de "décalage", qui peut être positif ou négatif. Ce type de dissemblance est peu dépendante du contenu linguistique du signal, si celui-ci est assez long. Elle

caractérise en premier lieu la voix de la locutrice et les conditions d'enregistrement.

La seconde s'apparente à la covariance et cumule les distances (hors décalage) prises par l'indice tout au long de chaque séquence. C'est une distance euclidienne. Nous parlerons de "dissemblance de profil", ou plus simplement "profil". Ce type de dissemblance caractérise le contenu linguistique et le style d'élocution.

Nous obtenons donc, pour évaluer les différences entre deux séquences, une série de 16 mesures regroupées en quatre groupes: décalage prosodique (Ddur, Dint, DF0, DdF0), décalage spectral (Dcc1, Dcc2, Dcc3, Dcc4), profil prosodique (Pdur, Pint, PF0, PdF0) et profil spectral (Pcc1, Pcc2, Pcc3, Pcc4).

Pour chaque mesure ou combinaison de mesures on peut classer les 96 séquences par distance croissante par rapport à la séquence moyenne. Il s'agit alors d'une distance euclidienne; pour les mesures de type "décalage" le signe de la différence à la moyenne n'est pas pris en compte

Il apparaît en premier lieu que la séquence 07 qui a servi de référence pour l'alignement n'est jamais, pour aucune mesure ou combinaison de mesures, la séquence la plus proche de la moyenne. Ceci valide a posteriori la normalisation temporelle par alignement décrite plus haut (§ 2.5), puisque la séquence de référence n'impose pas sa propre organisation prosodique au processus de comparaison.

En second lieu on peut tenter d'évaluer la qualité de chaque mesure ou combinaison de mesures en comparant les rangs (de 1 à 96) pris dans chaque classement par les deux séquences issues d'une même locutrice. L'idée directrice est que la voix et l'élocution restent sensiblement identiques d'une lecture à l'autre, et qu'une mesure est d'autant plus représentative de la réalité qu'elle classe les deux séquences à des rangs voisins. Le critère de cohérence retenu est la différence de rang, moyennée sur les 48 locutrices. Idéalement ce critère devrait être égal à 1, ce qui serait le cas pour deux séquences identiques. Le seuil de chance est de 48. En fait les valeurs observées varient entre 10 et 35 et sont donc assez loin de la valeur idéale de 1, pour plusieurs raisons: les deux séquences ne sont jamais parfaitement identiques (à cause des variations de prononciation et des défauts tels que reprises et hésitations); les mesures faites sur le signal sont entachées d'erreur (F0, durées); et les traitements effectués peuvent eux-mêmes engendrer des erreurs (sauts de GV lors de l'alignement).

Les résultats reportés dans le tableau 1 pour chaque indice montrent que la mesure DdF0 est à écarter, car sa moyenne est sensiblement nulle et les classements qu'elle engendre sont de mauvaise qualité. Par contre PdF0 montre une cohérence comparable à celle des autres mesures.

Le tableau 2 montre la cohérence obtenue pour chaque groupe de 4 mesures (à l'exception du groupe "décalage prosodique", dont Ddf0 est écarté). Il apparaît que, pris ensemble, les décalages conduisent à une meilleure cohérence que les profils. Une bonne valeur de cohérence (11,0) est obtenue avec un sous-ensemble de quatre mesures "décalage": Ddur, DF0, Dcc2, Dcc3. Une cohérence à peine moindre (12,6) est obtenue avec un ensemble de deux mesures "décalage" (DF0, Dcc2) et deux mesures "profil" (PF0, Pcc4).

TAB. 1 – *Cohérence obtenue avec chaque mesure prise isolément*

Ddur	Dint	DF0	DdF0	Dcc1	Dcc2	Dcc3	Dcc4
16,1	17,8	13,6	33,5	16,5	12,1	13,2	14,5

Pdur	Pint	PF0	PdF0	Pcc1	Pcc2	Pcc3	Pcc4
17,8	24,8	14,7	23,3	24,5	16,3	20,8	16,0

TAB. 2 – *Cohérence obtenue avec chaque groupe de 4 mesures (3 dans le premier groupe)*

décalage prosodique	décalage spectral	profil prosodique	profil spectral
13,6	13,8	21,0	17,5

Ces évaluations confirment la relative fiabilité des indices et mesures que nous avons définis, pour comparer les séquences les unes aux autres.

4.2 Exemples de variations prosodiques

La figure 2 montre l'évolution de F0 pour la séquence 52, classée sur l'ensemble des mesures comme la plus proche de la moyenne, et la séquence 40, classée comme la plus éloignée. On observe en effet de nombreuses différences, même si le profil d'ensemble est comparable. En premier lieu la voix est plus aiguë (d'environ 5 demi-tons), et l'étendue des variations est plus grande (même en échelle logarithmique). A l'écoute cette voix peut être qualifiée de rapide, précise, enfantine, chantante. En second lieu le détail du profil de F0 montre des déplacements des maxima de F0 qui traduisent une interprétation différente de certaines phrases.

La figure 3 montre l'évolution de F0 pour la séquence 52 et pour la séquence 79, classée comme proche de la moyenne pour les trois mesures de profil prosodique. On constate en effet une très grande parenté entre ces séquences, non seulement dans la moyenne et l'étendue de F0, mais aussi dans la position des extréma de F0. Comme pour la locutrice 40, ces déplacements d'accent ne sont pas nécessairement des erreurs. Le texte laisse la lectrice libre d'accentuer comme elle l'entend certains mots ou phrases, et de faire passer ainsi une nuance personnelle dans son élocution.

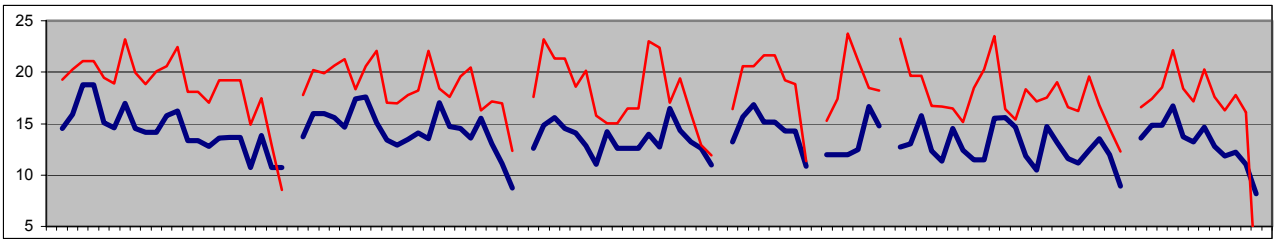

FIG. 2 – *Evolution de F0 (demi-tons) pour les séquences 52 (la plus proche de la moyenne, toutes mesures confondues - en trait épais), et 40 (la plus éloignée - en trait fin)*

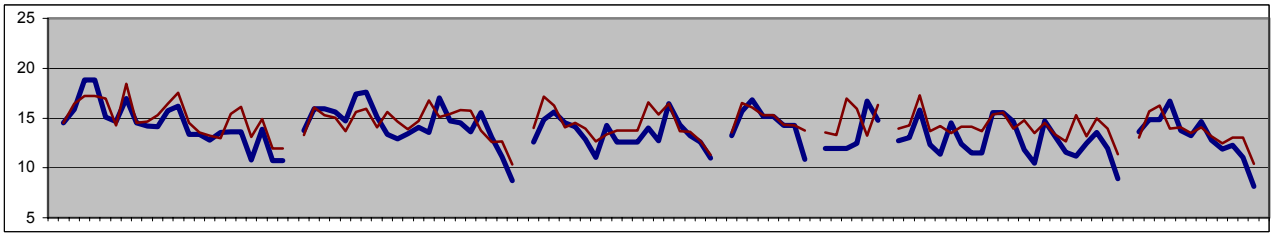

FIG. 3 – *Evolution de F0 (demi-tons) pour les séquences 52 (trait épais) et 79 (la plus proche de la moyenne pour les mesures de profil prosodique - en trait fin)*

5. CONCLUSION

A partir d'un corpus de parole lue dans des conditions homogènes nous avons pu mettre en évidence une prosodie moyenne pour ce style d'élocution, respectée dans ses grandes lignes par toutes les locutrices.

La moyenne fait apparaître divers groupements hiérarchiques: groupes interpausaux imposés par les ponctuations du texte, mots prosodiques dont la structure semble liée à la langue et, entre les deux, groupes dont la structure est liée à l'organisation, au sens et à l'interprétation de la phrase.

Les variations reflètent des différences imputables à la locutrice en tant qu'individu physique, ainsi qu'à sa manière de parler. Nous avons défini deux types de distances pour chaque indice, l'une appelée décalage (valeurs moyennes, reflétant plutôt les caractéristiques vocales de la locutrice indépendamment du texte et du style), l'autre appelée profil et plutôt liée à sa manière de parler particulière. La qualité de ces mesures a été évaluée par la cohérence des classements qu'elles fournissent pour deux séquences de la même locutrice. On peut ainsi dire en quoi, de combien, et où la prosodie d'une séquence donnée s'écarte de la moyenne.

Deux points au moins méritent d'être approfondis dans la suite de cette étude. D'une part, il semble que le découpage du texte par les pauses, dont certaines seulement sont obligées, laisse une grande marge d'initiative à la locutrice. Ce découpage est essentiel: il détermine la structure prosodique du plus haut niveau. et engendre la forme la plus apparente du discours. D'autre part nous avons observé une contribution importante des paramètres spectraux (les premiers mfcc, qui représentent l'enveloppe fortement lissée du spectre) à la caractérisation des variations prosodiques. Il conviendrait de définir plus précisément en quoi ces indices sont liés à la prosodie individuelle du locuteur.

L'approche proposée ne repose pas sur la connaissance préalable du contenu linguistique; elle requiert seulement la certitude que celui-ci est le même d'une séquence à l'autre. Elle tente de séparer au mieux ce qui relève de la norme de prononciation d'un groupe donné de locuteurs, de ce qui est propre à chaque locuteur ou à chaque réalisation individuelle de la norme. Elle devrait être utilisable dans tous les cas où l'on dispose de plusieurs séquences de même contenu linguistique, présentant entre elles des différences prosodiques significatives.

RÉFÉRENCES

[1] J.S. Liénard "From speech variability to pattern processing": a non-reductive view of speech processing", in *Levels in Speech Communication : relations and interactions*, eds J.Schoentgen, et al. Elseviers Science Publishers, 1995

[2] P. Léon Précis de phonostylistique; parole et expressivité, ed. Fernand Nathan, 1993

[3] M.A. Leblanc "Recherche de correspondances entre production écrite et production orale", thèse de l'Univ. Paris X Nanterre, Nov. 2001

[4] P. Boersma and D. Weenink "PRAAT: doing phonetics by computer" http://www.fon.hum.uva.nl/praat/

[5] J.B. Pierrehumbert "Synthesizing intonation" J. Acoust. Soc. Am. 70, 985-995, 1981

[6] N. Grönnum "The Groundworks of Danish Intonation", Copenhagen: Museum Tusculanum Press, 1992

[7] A. Botinis *Intonation: Analysis, Modelling and Technology*, Kluwer Ac. Publ., Dordrecht, 2000