Vers un algorithme de détection (semi-) automatique des proéminences en français parlé

Jean-Philippe Goldman, Mathieu Avanzi

Vers une automatisation de la détection des proéminences en français parlé

Jean-Philippe Goldman[1] ; Mathieu Avanzi[2]

[1] Département de Linguistique, Faculté des Lettres, Université de Genève, CH-1211 Genève, 4, Suisse
[2] Institut de philologie romane et linguistique française, Université de Neuchâtel, CH-2000 Neuchâtel, Suisse
goldman@lettres.unige.ch ; mathieu.avanzi@unine.ch

ABSTRACT

Prosodic transcription of spoken corpora relies mainly on the identification of perceived prominence. However, the manual annotation of prominent phenomena is extremely time-consuming, and varies greatly from one expert to another. Automating this procedure would be of great importance. In this study, we present the first results of a methodology aiming at an automatic detection of prominence syllables. It is based on (i) a spontaneous French corpus that has been manually annotated according to a strict methodology and (ii) some acoustic prosodic parameters, shown to be corpus-independent, that are used to detect prominent syllables. Some automatic tools, used to handle large corpora, are also described.

1. INTRODUCTION

S'il est important que la communauté scientifique dispose de corpus de langue parlée annotés prosodiquement, il est aussi important que ces annotations soient consensuelles, afin de faciliter les échanges et les comparaisons. Le point d'ancrage de telles annotations s'articulent en général autour du repérage de proéminences : c'est en fonction des proéminences détectées le long de la chaîne parlée que l'on va générer et interpréter la structure prosodique. L'enjeu est d'autant plus capital que les problèmes associés à cette tâche sont aujourd'hui loin d'être résolus. Pour preuve, l'expérience de [Poi06] basée sur l'identification perceptive de proéminences par 7 experts, et dans laquelle un extrait d'environ 3 minutes de parole spontanée produite par un locuteur belge a été choisi pour réaliser la tâche d'identification des proéminences, en position finale et non finale. Parmi les 165 syllabes analysées, la variation dans le pourcentage des syllabes reconnues comme proéminentes est telle (de 19 % à 49 %) qu'il semble raisonnable de conclure que les sujets ne partagent pas la même définition du concept. Par la suite, [Mor06] ont conduit une analyse phonétique du corpus pour établir une corrélation quantifiée entre les objets acoustiques saillants (en termes de paramètres prosodiques et de seuils mobilisés de hauteur mélodique, intensité et durée) et l'annotation manuelle des proéminences. Les conclusions de l'analyse mettaient en lumière la complexité de ces corrélats et la nécessité d'en dresser un inventaire raisonné. Ils concluaient sur l'impossibilité d'envisager une détection robuste sans l'apport de logiciels de détection automatique. Ces logiciels doivent tenir compte non seulement des indices de F0 (caractéristique la mieux perçue), mais aussi de l'intensité et de la durée sous forme de débit local, paramètre assez bien représenté en français par la durée de chaque syllabe par rapport à la durée moyenne des syllabes environnantes Bel[06]. De même, l'intensité reste peu prise en compte alors que des tests en synthèse montrent qu'elle n'est pas négligeable du tout. Enfin, il paraît recommandé d'étudier ces trois caractéristiques simultanément, en particulier leurs corrélations. La détection automatique que nous proposons, dérive de ces différentes constatations, couplées aux hypothèses formulées en phonétique expérimentale sur les seuils perceptifs de l'accent [Har91] ; [Mer87]. En pratique, l'étiquetage accentuel repose non pas sur une propriété structurelle abstraite du mot ou du groupe de mots comparable à la notion de *(lexical) stress* [Mar06] mais sur une définition phonétique neutre de la notion de proéminence. Cette dernière est associée à une saillance perceptive sur un fond sonore. Une telle approche présente l'intérêt d'être largement consensuelle et indépendante des cadres théoriques envisagés. Afin d'implémenter automatiquement un algorithme de détection automatique des proéminences en français parlé, un corpus (section 2) a été annoté manuellement par deux experts codeurs. Ces derniers ont mis en place un protocole strict pour le codage des proéminences. Celui-ci est détaillé dans [Ava07] et résumé dans la section 3. L'outil que nous avons élaboré repose sur la prise en compte de paramètres acoustiques basiques. Dans la dernière partie de cet article, nous rappelons de quelle façon l'outil a été constitué, et les résultats qu'il rend possible.

2. CORPUS D'ÉTUDE

Le corpus sur lequel se base cette étude est constitué de deux types d'enregistrements, à dominante monologale, d'une durée totale de 20 minutes. Des prescriptions d'itinéraires (cote Iti) *in situ*, recueillis à micro ouvert dans la région grenobloise (7 extraits dont 2 femmes) ; des interviews radiophoniques des radios publiques française et belge (cote Irt, 2 extraits dont une femme).

Les transcriptions en orthographe standard ont été phonétisées puis alignées semi-automatiquement sur les sons à l'aide de l'outil *EasyAlign* [Gol07], qui fonctionne sous *Praat* [Boe07].

3. DÉTECTION AUDITIVE DES PROÉMINENCES

Dans un premier temps, deux experts en prosodie ont annoté les proéminences syllabiques du corpus selon un protocole strict. L'annotation de corpus de français parlé non préparé pose des problèmes que les études réalisées

sur des corpus de parole lue et contrôlée ne rencontrent pas. Les phénomènes typiques de l'oral et liés à la production, comme les hésitations, les faux départs ou les interruptions, doivent être traités de manière spécifique pour éviter qu'ils interfèrent dans la fiabilité de l'annotation/détection des proéminences. Nous avons retenu deux ensembles de symboles pour le codage (cf. tableau 1).

3.1 L'étiquetage des proéminences proprement dites se fait avec les symboles P, p et 0. La distinction entre « proéminence forte » et « proéminence faible » a une fonction heuristique : elle force à développer une écoute plus fine. Lors de l'analyse, les syllabes « P » et « p » sont regroupées en une catégorie, qui s'oppose à « 0 ».

3.2 L'étiquetage de phénomènes typiques de la production d'oral spontané fait l'objet d'une tire d'exception (***delivery***) : elle permet d'isoler des syllabes que chaque codeur, selon ses a priori théoriques, pourrait coder d'une manière propre, engendrant par là des divergences ou des incohérences quant à la perception des proéminences. Ces syllabes (**hésitations, interruptions, schwas post-toniques**) peuvent faire l'objet d'un traitement spécifique ultérieur. Les parties impossibles à exploiter acoustiquement, et parfois même à segmenter en syllabes (**rires, toux, chevauchements, etc**.) sont notées « % » et exclues. Les silences « _ » sont détectés automatiquement lors de l'alignement ; les **prises de souffle** « * » sont notées manuellement.

La liste des symboles employés pour le codage du corpus est dressée dans le tableau 1.

Table 1. Symboles pour l'annotation

1. Codage des proéminences	
P	Proéminence forte
p	Proéminence faible
0	Syllabe non proéminente
2. Codage des phénomènes de *delivery*	
z	Hésitation (allongement,euh,creaky voice)
@	Schwa postonique (comme dans « c'est dinguE », [sEde~g@]
$	Appendice (syllabe(s) postoniques non accentuées)
!	Interruption de mot ou de syntagme
%	Partie inaudible ou inexploitable (rire, toux, chevauchement, bruit)
*****	Prise de souffle
_	Silence (issu de la détection automatique)

La procédure de codage va comme suit. Les deux annotateurs écoutent de courts extraits du fichier audio (d'une durée approximative de 3,5 sec.), et le rejouent jusqu'à trois fois. Les syllabes perçues proéminentes sont codées « p » ou « P ». Ils font de même avec les autres symboles de delivery. Une syllabe peut être codée à la fois comme proéminente et appartenant à une amorce de constituant syntaxique (p ! ou P !), mais une syllabe ne peut pas être codée comme à la fois proéminente et faisant partie d'une hésitation. Ceci constitue une des rares, sinon la seule, contraintes « théoriques » de notre procédure. Les deux tires qui en résultent ont ensuite été comparées puis les désaccords ont été discutés et réglés au cours de discussions communes. Il en ressort une annotation consensuelle, dont le tableau 2 donne les résultats :

Table 2. Statistiques finales de l'annotation manuelle sur le corpus d'étude. De bas en haut : durée des corpus et nombre de syllabes total ; syllabes exclues (marquées par un signe de delivery) – avec pourcentage par rapport à la totalité des syllabes ; syllabes proéminentes, syllabes non proéminentes – avec pourcentage par rapport à la totalité des syllabes ; total de syllabes « valides », *i.e.* ensemble par rapport auquel doit se mesurer l'automate

	Irt-LF	Irt-WL	Iti-10	Iti-12	Iti-14	Iti-22	Iti-B	Iti-D	Iti-S	**Total**
durée (s)	331	295	50	46	100	203	27	128	33	**1213**
tot.syll	1403	1195	181	148	430	820	128	436	140	**4881**
delivery	124	141	20	24	65	73	12	37	14	**509**
%	8,83	11,79	11,04	16,21	15,11	8,90	9,37	8,48	10	**10,42**
P/p	333	314	35	42	104	192	27	106	30	**1182**
%	23,73	26,27	18,33	28,37	24,18	23,41	21,09	24,31	21,42	**24,21**
non P/p	946	740	126	82	261	555	89	293	96	**3190**
%	67,42	61,92	69,61	55,40	60,69	67,68	69,53	67,20	68,57	**65,35**
syll.valid.	**1279**	**1054**	**161**	**124**	**365**	**747**	**116**	**399**	**126**	**4372**

Pour information, le pourcentage d'accord inter-annotateur atteint 89,35 % [Ava07]. On peut considérer que ce score comme vraiment satisfaisant, puisque d'après les recensions faites par [Tam05 : 43], le taux d'accord entre deux experts dans l'identification des proéminences prosodiques oscille généralement entre 80 et 84 %, dans le meilleur des cas. Signalons aussi que l'annotation peut prendre 5 fois le temps réel du corpus à coder, une fois que celui-ci est aligné en syllabes.

Au final, sur un total de 4881 intervalles syllabiques, 509 syllabes ont été exclues via la tire d'exception (10,42%), 1182 syllabes ont été codées p ou P (soit 24,21%). Restent 3190 syllabes non proéminentes (65,35 %). L'outil de détection automatique mis en place prendra comme mesure de référence le nombre total de syllabes non exclues, qui s'élève à 4372 unités.

4. DÉTECTION AUTOMATIQUE DES PROÉMINENCES

Les résultats de cette identification auditive experte servent de référence pour l'entraînement et la validation d'un système automatique de détection de syllabes proéminentes, basé sur des paramètres prosodiques comme la mélodie, l'intensité et la durée. La procédure se déroule en deux temps : 1. identification des noyaux vocaliques; 2. mise en rapport des paramètres acoustiques prosodiques pour la détection de proéminences.

La première étape prend pour point de départ un script élaboré par [Mer04] fonctionnant avec le logiciel Praat : le Prosogramme. A l'origine, cet outil avait été développé pour faciliter la transcription semi-automatique de la prosodie, en opérant à une stylisation de la mélodie. Cette

stylisation peut être faite à partir du signal seul, mais elle est plus robuste si un alignement phonétique est fourni. Pour chaque syllabe, le noyau vocalique est délimité comme la partie voisée qui présente une intensité suffisante (en utilisant des seuils relatifs au pic d'intensité local). Puis, pour chaque noyau, la F0 est stylisée en un ou plusieurs segments de droite. Ces segments peuvent être stylisés comme plats ou avec une pente mélodique, selon des seuils perceptuels de glissando qui sont réglables.

Malheureusement, dans nos corpus, un nombre non négligeable de noyaux n'a pas été détecté ou mal stylisé. Les raisons en sont diverses : en premier lieu, la version originale du Prosogramme a été développée pour styliser de la parole avec ou sans l'aide d'une segmentation phonétique préalable. De ce fait, certains seuils non réglables par l'utilisateur ont été ajustés de la même manière pour les deux modes de fonctionnement. La seconde source d'erreurs concerne les frontières de phonèmes. Celles-ci sont détectées par la segmentation avec EasyAlign [Gol07] mais il est possible qu'elles puissent ne pas être tout à fait exactes. Enfin, le paramètre d'intensité utilisé pour la segmentation des noyaux vocaliques n'est pas entièrement fiable, parce qu'il est à la fois instable et dépendant de la nature des segments. En effet dans la version originale du Prosogramme, si le pic d'intensité a lieu dans une partie non voisée de la consonne, aucun noyau n'est détecté.

Pour pallier ces problèmes, et, afin qu'un maximum de noyaux soient stylisés, nous avons apporté de légères modifications à la version originale du Prosogramme : 1. Si le pic d'intensité, utilisé comme « base de construction du noyau syllabique », est recherché dans la voyelle (délimitée par la segmentation phonétique) et qu'il n'y est pas trouvé, on autorise la stylisation à s'étendre à l'attaque et/ou à la coda de la syllabe complète, tout en maintenant la contrainte de voisement. 2. Des systèmes de routines de *back-off* ont été implémentés pour forcer la détection du noyau. 3. Nous avons rendu réglables certains paramètres comme les seuils d'intensité pour la segmentation du noyau.

Ces modifications se sont avéré fort utiles, puisqu'au final, la proportion de noyaux stylisés est passée de 85% à 95%.

Une fois les noyaux détectés, nous avons comptabilisé pour chacun des intervalles syllabiques les mesures acoustiques suivantes :

1. La durée de la syllabe (en millisecondes), que l'on préfère intuitivement à la durée du noyau syllabique stylisé, laquelle est dépendante des traits de voisement des consonnes contenues dans l'onset et la coda ;
2. Le moyenne de F0 sur le noyau (en semi-tons)

Ces premières mesures acoustiques sont ensuite « relativisées », i.e. recalculées par rapport au contexte syllabique immédiat (par rapport au deux syllabes précédentes et à la syllabe suivante), pour obtenir des durées relatives (sans unités), et des mesures de F0 relatives (en ST). Les pauses silencieuses de plus de 250 ms ont été utilisées pour contraindre ce calcul : certaines de ces syllabes adjacentes, normalement utilisées pour le calcul des paramètres relatifs, peuvent être ignorées si elles sont au-delà d'une pause par rapport à la syllabe en cours de calcul.

5. RÉSULTATS

Sur la base de ces paramètres, une stratégie de décision basée sur des seuils acoustiques a été testée sur les 4372 syllabes valides (c'est-à-dire non exclues par la tire *delivery*). Il s'agit de déterminer la pertinence des deux paramètres prosodiques relatifs pour estimer le caractère proéminent ou non des syllabes. En prenant comme référence l'annotation manuelle, **84.54%** des syllabes sont correctement reconnues. Une syllabe sera considérée comme proéminente si l'un ou l'autre des deux paramètres (F0 max ou durée syllabique) est supérieur à un seuil. Seront donc considérées comme proéminentes les syllabes hautes ou/et longues relativement aux syllabes adjacentes. Les seuils optimaux pour ce corpus sont : **3 ST pour le paramètre de hauteur et 2 comme durée syllabique relative**

Parmi ces syllabes « bien classifiées », 66.7% sont des syllabes identifiées comme non proéminentes par les humains et détectées comme telles par l'automate, et 17.8 % sont des syllabes identifiées comme proéminentes par l'annotation manuelle et l'automate. Le reste sont des fausses alertes (5.5 % de syllabes saillantes uniquement selon l'algorithme) ou des détections manquées (10% des syllabes marquées P par les experts ne sont pas identifiées comme telles par l'automate).

Le Prosogramme original permet de tracer des graphiques représentant la mélodie stylisée en regard de la segmentation phonétique. La version modifiée présentée ici y ajoute quelques informations, comme : 1.La valeur des paramètres prosodiques relatifs pour chaque noyau (en vert et souscrit, la durée relative, en rouge et suscrit la hauteur mélodique relative), et 2. la tire d'annotation manuelle, comprenant à la fois les proéminences mais aussi les symboles de la tire *delivery* comme les hésitations, les faux départs… (en bleu au-dessus de chaque segment stylisé). Les syllabes annotées proéminentes par les experts et identifiées comme telles par l'algorithme sont colorées en rouge, les syllabes annotées proéminentes par les experts mais non reconnues comme telles par l'algorithme sont en gris (détection manquée) ; enfin, dernier cas de figure, les fausses alertes (syllabes proéminentes selon l'automate mais non codées P/p par les experts) sont en magenta.

Cette façon de faire permet à l'utilisateur de travailler sur une version du corpus dans laquelle le discours est réduit à ses seules informations pertinentes du point de vue des proéminences, à savoir, selon notre hypothèse : les pics conjoints d'intensité et de F0, la stylisation de la fondamentale, la segmentation phonétique et syllabique et les valeurs des paramètres acoustiques retenus. Chaque erreur de détection peut ainsi être diagnostiquée *a posteriori* (mauvaise segmentation, codage incertain, non pertinence des paramètres acoustiques choisis, ...).

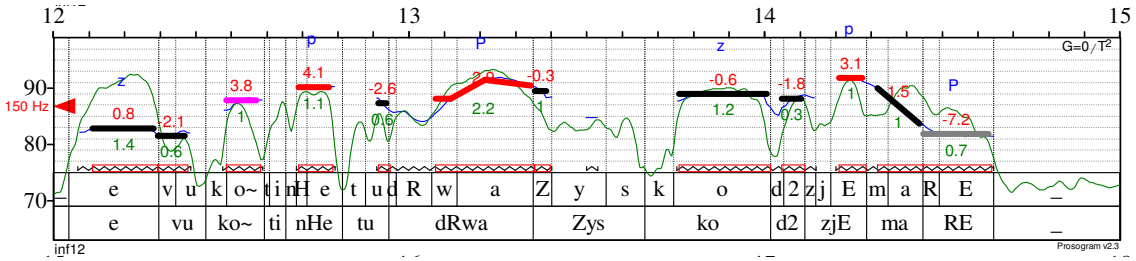

Figure 1: Prosogramme enrichi (Iti-12) « *et vous continuez tout droit jusqu'au deuxième arrêt.* » L'intensité et la courbes de F0 effectivement stylisé sont en traits gras noirs. Les valeurs des paramètres acoustiques et le codage manuel apparaissent en dessous et au-dessus des noyaux stylisés (avec, de bas en haut : durée relative des syllabes, hauteur de F0 relative, codage manuel.

5. CONCLUSION

Dans cet article, nous avons fait état une approche à deux volets devant aboutir à l'annotation semi-automatique des proéminences syllabiques dans des corpus oraux non lus. Cette entreprise se justifie par le fait que les proéminences sont des phénomènes acoustiques basiques, préliminaires obligés à toute étude s'intéressant aux rapports entre prosodie et discours. L'outillage présenté et discuté ici se montre à la fois souple et robuste. Il permet d'envisager de traiter automatiquement ou semi-automatiquement de grandes masses de données orales, et d'extraire des résultats statistiques dont la validité est fonction de l'ampleur du corpus, et de la fiabilité de l'outil automatique. Sa valeur, son utilité réelle, est fonction des tâches qui lui seront confiées et auxquelles il sera confronté. Outre une meilleure connaissance des prosodies, de la parole lue/préparée ***et*** de la parole spontanée, il peut servir à caractériser, en interaction avec une analyse linguistique et discursive, des stratégies prosodiques associées à des buts communicatifs, de façon plus ou moins locale, aussi bien que des profils phonostylistiques, récurrents ou occasionnels, associés à des sociolectes, dialectes, etc. – c'est-à-dire aussi bien à repérer ou détecter des points communs et des différences entres variantes sélectionnées qu'à décrire avec précision les caractéristiques de telle ou telle variante [Sim07]. Une fois les prolongements annoncés effectués, nous espérons que cet outil pourra être utilisé pour l'annotation des grands corpus de langue parlée

6. REMERCIEMENTS

Ce travail a reçu le soutien financé du FNS (subside subside n°100012-113726/1). Les auteurs remercient également A. Auchlin, A. Lacheret-Dujour, A-C. Simon qui les ont autorisés à réexploiter les résultats présentés dans cette étude, qui s'inscrit dans le cadre d'un travail plus large sur la constitution d'un corpus de référence annoté prosodiquement, et d'outils pour l'exploiter.

RÉFÉRENCES

[Ava07] Avanzi, M., Goldman, J.-Ph., Lacheret-Dujour, A. et Simon, A-.C. (2007), « Méthodologie et algorithmes pour la détection automatique des syllabes proéminentes dans les corpus de français parlé », à par. *Cahiers of French Language Studies*.

[Bel06] Beller, G. et al (2006), « Speech Rates in French Expressive Speech », *Proc of Speech Prosody*

[Boe07] Boersma, P. & Weenink, D. (2007), *Praat: doing phonetics by computer* www.praat.org

[Gol07] Goldman, J.-Ph. (2007), « EasyAlign : a semi-automatic phonetic alignement tool under Praat » http://latcui.unige.ch/phonetique

[Har91] t'Hart, J. Collier, R. & Cohen, A. (1991), *A Perceptual Study of Intonation. An Experimental-Phonetic Approach to Speech Melody*. Cambridge, University Press.

[Mar06] Martin, Ph. (2006), « La transcription des proéminences accentuelles : mission impossible ? », *Bulletin PFC*, 6, pp. 81-87.

[Mer87] Mertens, P. (1987), *L'intonation du français. De la description linguistique à la reconnaissance automatique*, Unpublished Ph.D. (Univ. Leuven, Belgium).

[Mer04] Mertens, P. (2004), « Un outil pour la transcription de la prosodie dans les corpus oraux », *Traitement Automatique des langues* 45 (2), pp. 109-130.

[Mor06] Morel, M., Lacheret-Dujour, A., Lyche, C. & Poiré, F. (2006), « Vous avez dit proéminences ? », *JEP 06*, pp. 183-186.

[Poi06] Poiré, P. (2006), « La perception des proéminences et le codage prosodique », *Bulletin PFC* 6, pp. 69-79.

[Sim07] Simon, A.C, Auchlin, A. Goldman, J.P. & Avanzi, M. (2007), *Nouveaux Cahiers de Linguistique Française,* Actes du colloque IDP07, 12-14 septembre 2007, Genève.

[Tam05] Tamburini, F. & Caini, C. (2005), « An Automatic System for Detecting Prosodic Prominence in American English Continuous Speech », *International Journal of Speech Technology* 8, pp. 33-44.