Comparaison de trois outils de détection automatique de proéminences en français parlé

Anne Lacheret, Nicolas Obin, Jean-Philippe Goldman, Mathieu Avanzi

Comparaison de trois outils de détection automatique de proéminences en français parlé

[1]N. Obin, ,[2,5]J.-Ph. Goldman, [3]M. Avanzi, [4]A. Lacheret
[1]IRCAM, [2]Université de Neuchâtel, [3]Université de Genève,
[4]Université de Paris X, MODYCO Nanterre et IUF, [5]Université de Louvain-la-Neuve
Nicolas.Obin@ircam.fr; Jean-Philippe.Godman@lettres.unige.ch; Mathieu.Avanzi@unine.ch; anne@lacheret.com

ABSTRACT

This paper presents the inner details of three different algorithms for prominence detection. On the basis of a 50-minute corpus made of five speaking styles and manually annotated for prominence, a quantitative evaluation compares the three approaches.

Keywords: spontaneous speech, prosody, prominence, automatic detection

1. INTRODUCTION

L'annotation des proéminences accentuelles qui scandent le discours spontané ne peut plus être envisagée manuellement, étant donné d'une part l'aspect subjectif de la démarche, d'autre part le temps de codage demandé, d'autant plus couteux qu'on souhaite aboutir à une annotation rigoureuse et stabilisée à l'issue de la confrontation de plusieurs expertises de codage. Une telle pratique manuelle est d'autant moins concevable qu'il s'agit de brasser des données de plus en plus volumineuses, quantitativement représentatives pour pouvoir les exploiter de manière fiable sous l'angle de l'analyse structurale et fonctionnelle. En pratique, les méthodes de traitement automatiques se développent [1] et doivent continuer à se développer pour prendre le relais du codage manuel et, tout en capitalisant les connaissances acquises, les faire évoluer et enrichir les analyses prosodiques de la parole non lue, des discours formels aux discussions à bâtons rompus. Cette communication s'inscrit dans cette problématique. Un corpus de français parlé échantillonné en différents genres, synthétiquement présenté dans la section 2, est utilisé pour comparer trois outils de détection automatique de proéminences, exposés dans la section 3. La section 4 fait état du protocole d'évaluation et la partie 5 présente les résultats.

2. MATÉRIEL D'ÉTUDE

2.1. Le corpus

Pour cette étude, nous avons utilisé un corpus échantillonné en cinq différents genres et styles de parole (discours politique – cote DP, descriptions d'itinéraires – cote ITI, récits de vie – cote RCV, journaux radiophoniques – cote JP et interviews radiophoniques, cote IRT). Les locuteurs sont tous des francophones natifs, et proviennent de France, de Suisse ou de Belgique). On en trouvera une présentation exhaustive dans [2].

2.2. Prétraitement et transcription

Le corpus a été transcrit et aligné semi-automatiquement (phonèmes, syllabes et mots graphémiques) sous Praat [3] avec EasyAlign [4]. Les alignements ont fait l'objet d'une vérification manuelle par deux experts humains, qui ont ensuite consigné dans une tire dédiée (tire *delivery*) les informations relatives au statut des syllabes (syllabes proéminentes ou non ; syllabes particulières contenant des phénomènes propres au langage spontané, *i.e.* hésitations, faux-départ, schwas post-toniques, bruits de bouche, etc). On se reportera aux travaux de [5] pour un compte-rendu exhaustif sur la procédure et ses origines. Au total, le corpus est composé de 12851 intervalles syllabiques. 973 d'entre eux ont été exclus via la tire *delivery*, 3244 syllabes ont été annotées proéminentes, 8634 ont été codées ni proéminentes ni associées à un marqueur *delivery*, soit 11878 à traiter automatiquement.

3. LES LOGICIELS DE DÉTECTION

Dans cette partie, nous décrivons les principes de base de chacun des trois outils dont nous souhaitons comparer les performances sur une tâche de détection automatique de proéminences.

3.1. Analor

Analor est un logiciel d'analyse implémenté sous Matlab [6] qui fonctionne avec des fichiers xml en sortie de Praat. Initialement, le logiciel a été conçu dans le but de faire émerger des critères acoustiques robustes en vue de segmenter automatiquement le continuum sonore en unités d'intégration prosodique maximales, ou **périodes intonatives**. Un ensemble de fonctions récemment élaboré permet aujourd'hui une détection automatique des syllabes proéminentes à l'intérieur des périodes identifiées par le logiciel. Cette procédure a été présentée et expérimentée pour la première fois dans [7]. Elle repose sur la mise au jour des variations significatives de hauteur et de durée par rapport à la moyenne globale de l'ensemble des syllabes qui composent une période.

Appelons $M_h(P)$ la moyenne et $E_h(P)$ l'écart-type de la fondamentale F0 sur une période P. Une syllabe s de P est dite proéminente pour la hauteur si elle contient un maximum local de la F0, noté h(s), vérifiant la condition :

$$\mathbf{h(s) > M_h(P) + K_h * E_h(P)}$$

où K_h est un paramètre ajustable indépendant du locuteur fixé à 1,5 par défaut.

Le calcul est le même pour la durée (pour une justification de ce seuil *a priori*, voir [8] et [9]). Les pauses silencieuses ont été également utilisées pour affiner la détection des syllabes proéminentes : toute syllabe suivie d'une pause est identifiée comme proéminente, *i.e.* saillante perceptivement.

3.2. ProsoProm

La détection automatique des proéminences dans par l'outil ProsoProm se déroule en trois étapes successives : 1.segmentation et stylisation des noyaux vocaliques, 2.extraction et relativisation des paramètres acoustiques et 3.décision du statut proéminent pour chaque syllabe.

La première étape utilise une version adaptée du ProsoGram de Mertens [10]. Cet outil, développé initialement pour de la transcription prosodique semi-automatique, repère et stylise le noyau vocalique de chaque syllabe. Plus précisément, dans chaque syllabe, le noyau est délimité comme la portion voisée qui « a suffisamment d'intensité » (en se basant sur des seuils d'intensité relativement au maximum d'intensité local). Cette étape permet d'éliminer des erreurs de détection de F0 aux frontières de voisement. Puis, la courbe mélodique de ce noyau est stylisée en un ou plusieurs segments, plats ou avec une pente mélodique, selon des paramètres perceptuels comme le glissando. ProsoGram peut repérer automatiquement les syllabes (et leur noyau) ou s'appuyer sur une segmentation phonétique. L'adaptation, décrite dans [11] permet d'augmenter le nombre de noyaux effectivement stylisés et d'affiner certains paramètres de réglages pour cette application spécifique.

Dans un second temps, plusieurs paramètres acoustiques sont estimés pour chaque syllabe, à savoir :

1. la durée syllabique, préférée à la durée du noyau car ce dernier est fortement contraint par le voisement de chacun des phonèmes composant la syllabe en secondes.
2. la hauteur mélodique maximale du noyau stylisé en demi-tons, considéré comme cible mélodique atteinte par le locuteur.
3. l'amplitude du mouvement montant en demi-tons. Le mouvement descendant n'est pas pris en considération dans l'idée qu'il est plutôt, à l'exception peut-être des syllabes finales, la manifestation d'un relâchement articulatoire.
4. la durée de la pause subséquente en secondes. Ceci restreint fortement notre étude à des langues oxytoniques comme le français.

Les deux premiers paramètres sont « relativisés » par rapport aux noyaux adjacents, c'est-à-dire que leur sont soustraits la moyenne des paramètres acoustiques correspondants des deux syllabes précédentes et de la syllabe suivante. Ceci permet de rendre ces paramètres indépendants du débit et du registre de parole. L'empan de relativisation choisi est volontairement très local rejoignant l'hypothèse que malgré l'existence évidente d'unités intonatives plus larges que quelques syllabes, l'appui sur les syllabes immédiatement adjacentes est primordial pour réaliser un effet de contraste.

Finalement, la stratégie de décision consiste à comparer chacun des paramètres avec un seuil et de décréter proéminente toute syllabe dont un des paramètres est au-dessus de son seuil propre. ProsoGram permet également de produire une sortie graphique dans laquelle les segmentations phonétiques et lexicales sont associées aux courbes mélodiques et d'intensité ainsi qu'au tracé des noyaux stylisés, dont la couleur dépend de leur caractère proéminent. Le fort pouvoir explicatif de ce graphique permet un diagnostic qualitatif des erreurs subsistantes.

Pour la présente étude, l'optimisation des quatre seuils pour chaque sous-corpus d'entrainement a été faite par une approche dichotomique.

3.3. IrcamProm

L'outil IrcamProm est un outil de détection de proéminences développé en Matlab dans le cadre d'IrcamCorpusTool [12]. Il repose sur un modèle statistique de la proéminence suivant un protocole décrit en [13]. Ce modèle se décompose schématiquement en deux étapes : 1. Une étape d'extraction et de sélection des paramètres acoustiques qui permettent la meilleure discrimination des syllabes proéminentes et non proéminentes. Les paramètres acoustiques utilisés dans la tâche de détection ont été sélectionnés automatiquement par un algorithme de sélection de descripteur à partir d'une description des paramètres acoustiques de la parole relativisés par rapport à plusieurs horizons temporels. 2. Une étape de modélisation à proprement dite. La modélisation est réalisée par un modèle de mélange de gaussiennes pour les syllabes proéminentes et non-proéminentes. Une fois les paramètres du modèle estimés, la décision de proéminence est réalisée suivant le critère de maximum *a posteriori*.

Les paramètres acoustiques retenus pour cette étude ont été obtenus par optimisation sur le corpus de parole lue présenté en [13]. Ils consistent en un jeu de paramètres acoustiques faisant intervenir plusieurs types d'information (hauteur, durée, information spectrale) relativisés sur plusieurs horizons temporels (absence, syllabes adjacentes et groupe prosodique). Nous les présentons dans l'ordre de leur importance relative pour la modélisation de la proéminence : la durée de la syllabe ; la valeur moyenne de la hauteur sur la syllabe relativisée par rapports à sa valeur moyenne sur les syllabes adjacentes ; la durée du noyau de la syllabe ; la valeur moyenne de la sonie dans la première bande de Bark sur la syllabe relativisée par rapport à sa valeur sur la syllabe suivante ; la valeur moyenne de la sonie dans la 18ème bande de Bark relativisée par rapport à sa valeur sur la syllabe précédente ; la valeur minimum du débit local sur la syllabe relativisée par rapport à la valeur minimum du débit local sur le groupe prosodique ; la courbure du débit local sur la syllabe.

Ce modèle a été élaboré sur un corpus de parole monolocuteur lue. Nous supposerons dans cette étude que ces paramètres estimés sur un corpus particulier demeurent inchangés pour le corpus présentement étudié, hypothèse qu'il restera à vérifier. En revanche, pour adapter notre modèle à une détection de proéminences sur de la parole spontanée multi-locuteur dans divers genres discursifs, nous avons normalisé l'ensemble des paramètres par rapport à leur valeur moyenne et déviation standard sur chaque groupe prosodique. Cette « relativisation » est pratiquée afin de normaliser l'influence du locuteur et du genre de discours sur la distribution des paramètres acoustiques.

4. PROTOCOLE D'ÉVALUATION

Les outils présentés ont été comparés sur le corpus décrit suivant un protocole de validation croisée. Cette méthode consiste à échantillonner le corpus en N parties non nécessairement de tailles égales. Alors les outils sont entraînés sur N-1 parties et validés sur la partie restante. Cette opération est répétée N fois en faisant varier le corpus de validation. Cette méthode présente deux avantages : 1. elle permet de tester la capacité de généralisation des outils en les testant sur des données non-observées au cours de l'apprentissage, et 2. les changements de corpus d'entraînement et de validation permet de tester la sensibilité de cette capacité par rapport au corpus d'entraînement.

Dans la présente étude, l'échantillonnage a été réalisé en découpant le corpus suivant les genres de discours, soit en cinq parties. Cela favorise l'indépendance de la partie d'entraînement et de validation. Comme l'outil Analor est un système de décision à base de règles, il n'était pas susceptible d'être entraîné. Nous avons donc utilisé les paramètres standard de cet outil directement sur l'ensemble des genres discursifs. Ce biais inséré est contrebalancé par le fait que les paramètres de cet outil n'ont pas été réglés sur le corpus que nous traitons présentement.

La mesure de performance des outils a été choisie comme étant la **f-mesure** de la classe proéminence, c'est-à-dire la moyenne harmonique des scores de précision et de rappel. Ce choix permet de faire la synthèse des performances en terme d'insertion et de délétion de proéminence.

La comparaison de la performance des outils est suivie d'une analyse automatique des différences de comportement observées entre ces outils. Cette analyse est fondée sur le test de MacNemar [14] qui permet de comparer statistiquement les différences observées sur les erreurs de classification entre les outils. Le test de l'hypothèse nulle est réalisé sur la base de ces différences pour déterminer si elles sont dûes ou non au hasard.

5. RÉSULTATS ET DISCUSSION

Nous présentons dans la Table 1 les performances des outils sur chaque genre de discours.

Table 1 : F-mesure des outils de détection sur chaque genre de discours **du corpus d'étude**.

	iti	irt	rcv	dp	jp	Total
Analor	68,7	71,5	63,2	74,5	70,8	**69,7**
ProsoProm	72,4	72,1	71,0	70,6	72,5	**71,7**
IrcamProm	75,8	74,3	76,3	76,00	75,0	**75,4**

Nous voyons de manière générale que les performances se situent aux environs de 70% de f-mesure en s'échelonnant de 70% à 75%. L'outil IrcamProm est l'outil qui présente la meilleure performance globale et par genre de discours. Analor apparaît plus sensible au genre de discours que les deux autres outils, avec une déviation standard relative six fois plus élevée que ProsoProm et IrcamProm. Ceci montre les limites de capacité de généralisation d'un système à base de règles.

Si nous approfondissons l'analyse des performances, nous voyons des tendances analogues pour l'ensemble des outils : ainsi, les trois outils présentent un minimum de performance sur un fichier de description d'itinéraire. Cette propriété suggère : 1. que la structure d'une proéminence diffère au moins quantitativement selon les genres de discours, et/ou 2. que le contraste proéminence/non proéminence est plus ou moins marqué selon les genres de discours.

Un diagnostic des erreurs montrent des comportements différents suivant les outils : Analor présente une tendance à la délétion de proéminence (rappel = 63.6%, précision = 77.2%) alors que ProsoProm présente une tendance à l'insertion de proéminence (rappel = 75.7%, précision = 68.2%). IrcamProm présente un compromis entre délétion et insertion (rappel = 76.4%, précision = 74.5%).

Une analyse du test de MacNemar ($p = 0.005$) sur l'ensemble des erreurs met au jour une différence significative entre l'outil IrcamProm et les deux autres outils, mais pas entre ces deux derniers. Néanmoins, cette analyse doit être relativisée par le fait que cette différence varie selon les genres de discours (nous observons par exemple une absence de diagnostic de différence sur les corpus d'itinéraires et de journaux parlés).

6. CONCLUSION

Nous avons exposé les performances de trois outils de détection automatique de proéminences dans un contexte de traitement de corpus de français parlé. Ces outils présentent chacun des particularités qui expliquent sans doute les scores obtenus et qui nous amènent à nous interroger et à mieux préciser l'impact des principes sous-jacents à la détection dans chaque système. Le premier, Analor, est un système à base de règles, alors que les deux autres sont fondés sur l'apprentissage. ProsoProm repose sur une stylisation perceptive de la courbe mélodique alors qu'Analor et IrcamProm se fondent sur des paramètres acoustiques bruts. IrcamProm prend en compte un nombre important de paramètres, alors que les deux autres algorithmes se focalisent sur la durée et les variations de F0, au mieux l'intensité (ProsoProm).

Dernière différence : Analor travaille sur une fenêtre périodique tandis que ProsoProm et IrcamProm n'intègrent que les contextes syllabiques immédiats.

A partir de ces constats, plusieurs questions se posent. Elles devront être prises en compte dans la suite de ce travail, à savoir le diagnostic quantifié des erreurs, en particulier :
- Peut-on mesurer l'impact d'une méthode fondée sur la stylisation perceptive (ProsoProm) par rapport à des approches travaillant uniquement sur la courbe acoustique brute ?
- Comment expliquer le paradoxe apparent : pourquoi le système à bases de règles (Analor) est plus sensible aux variations de genre discursif que les autres ?
- Quel est le meilleur compromis à trouver entre les résultats obtenus (performance de l'outil) et la complexité algorithmique demandée (voir IrcamProm) ?
- Quel est la meilleure fenêtre à prendre en compte dans un système par apprentissage ? Dans quelle mesure, dans un tel système, la période comme fenêtre d'analyse améliore-t-elle ou non les résultats par rapport à la seule prise en compte des syllabes immédiatement adjacentes ?

Enfin, rappelons que l'évaluation des outils s'est déroulée sur des syllabes non-marquées *delivery*, laissant ainsi de côté ces phénomènes propres à l'oral, et pour lesquels il est désormais possible d'envisager une détection automatique. Voilà l'ensemble des questions qui articulent notre programme de recherche à venir.

7. REMERCIEMENTS

Trois cadres institutionnels portent ce projet :

- Fonds National de la recherche scientifique Suisse (subside n°100012-113726/1, "La structure interne des périodes", Université de Neuchâtel),
- ANR Rhapsodie 07 Corp-030-01, Corpus prosodique de référence du français parlé.
- Programme Wist2 Convention n°616422, financé par la Région wallonne (Belgique) *EXPRESSIVE : Système automatique de diffusion vocale d'informations dédicacées : synthèse de la parole expressive à partir de textes balisés*

Nous souhaitons remercier ici A.-C. Simon et F. Poiré initiateurs de la méthode de codage manuel utilisée ainsi que B. Victorri pour sa collaboration active dans les différentes étapes de la modélisation.

BIBLIOGRAPHIE

[1] F. Tamburini, & C. Caini, An Automatic System for Detecting Prosodic Prominence in American English Continuous, Speech *International Journal of Speech Technology* 8: 33-44, 2005.

[2] A.-C. Simon, M. Avanzi & J.-P. Goldman. La détection des proéminences syllabiques. Un aller-retour entre l'annotation manuelle et le traitement automatique. Article soumis au *1er Congrès Mondial de Linguistique Française,* 2008.

[3] P. Boersma & D. Weenink. Praat: doing phonetics by computer (Version 4.5). www.praat.org, 2008.

[4] J.-Ph. Goldman. EasyAlign: a quasi-automatic phonetic alignment tool under Praat. *Journées d'Etudes sur la Parole*, Avignon, 2007 (soumis).

[5] M. Avanzi, J.P. Goldman, A. Lacheret-Dujour, A.C. Simon & A. Auchlin. Méthodologie et algorithmes pour la détection automatique des syllabes proéminentes dans les corpus de français parlé. *Cahiers of French Language Studies*, 13/2, 2007.

[6] A. Lacheret-Dujour & B. Victorri. La période intonative comme unité d'analyse pour l'étude du français parlé : modélisation prosodique et enjeux linguistiques. *Verbum, 24/1-2*. 55-73, 2002.

[7] M. Avanzi, A. Lacheret-Dujour & B. Victorri. ANALOR. A Tool for Semi-Automatic Annotation of French Prosodic Structure. In *Proceedings of Speech Prosody'08*, 2008.

[8] M. Rossi & al. : L'intonation, de l'acoustique à la sémantique, Paris, Klincksieck, 1981.

[9] A. Lacheret-Dujour & F. Beaugendre : La prosodie du français, Paris, CNRS éd., 1999.

[10] Mertens, P. The Prosogram: Semi-Automatic Transcription of Prosody based on a Tonal Perception" Model, in B. Bel & I. Marlien (eds.) Proceedings of Speech Prosody, Nara (Japan), 2004

[11] J.-P. Goldman, M. Avanzi, A. Lacheret-Dujour, A.-C. Simon & A. Auchlin. A Methodology for the Automatic Detection of Perceived Prominent Syllables in Spoken French. In *Proceedings of Interspeech'07*, pp. 91-120, 2007.

[12] C. Veaux, G. Beller, D. Schwarz, & X. Rodet, Ircamcorpustools: an extensible plateform for speech corpora exploitation, à paraître dans *The 6th edition of the Language Resources and Evaluation Conference*, Marrakech, 2008.

[13] N. Obin, X. Rodet & A. Lacheret-Dujour. French Prominence: a probabilistic framework. à paraître dans *International Conference on Audio, Speech and Signal Processing*, Las Vegas, Nevada, USA, 2008.

[14] T.G. Dietterich, Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms, in Neural Computation, Vol. 10, p. 1895-1923, 1998