CILL 30 (2004) n° 1-3, 7-25

LE PROSOGRAMME : UNE TRANSCRIPTION SEMI-AUTOMATIQUE DE LA PROSODIE[1]

Piet MERTENS
Université de Leuven

1. INTRODUCTION

Toute recherche linguistique sur l'intonation repose sur un travail descriptif qui passe inévitablement par la transcription prosodique d'énoncés ou, de préférence, d'un corpus de parole de taille appréciable. Depuis longtemps les syntacticiens de l'oral et les spécialistes de l'analyse du discours ont compris la nécessité de prendre en compte l'intonation. Toutefois, comme la transcription prosodique présuppose une compétence très spécialisée et représente un investissement en temps prohibitif, l'intégration de l'intonation était toujours remise aux calendes grecques. Le présent article décrit un système pour la transcription semi-automatique de la prosodie ; il sera question de sa conception et de son utilisation.

Quel genre de transcription adopter pour annoter un corpus ? Grosso modo on distingue trois types majeurs de représentations de la prosodie : l'analyse acoustique (paramètres de fréquence fondamentale, d'intensité et le spectre), la notation auditive et la notation symbolique de type phonologique ou tonologique. Ces trois types correspondent à des étapes (de traitement) dans la communication parlée. En effet, le signal sonore passe d'abord (et inévitablement) par un traitement auditif et perceptif,

[1] Cette recherche a été effectuée dans le cadre du projet plurifacultaire « Prosodie », subventionné par l'Université de Genève. Nous tenons à remercier J.-Ph. GOLDMAN pour l'annotation phonétique du corpus Groult, E. GEOFFROIS pour celle des corpus Barthes et Giroud (effectuée au LIMSI, Orsay, en 1994), A.-C. SIMON pour sa contribution à la transcription auditive manuelle du corpus Groult, et P. BOERSMA pour la mise à disposition du logiciel Praat.

avant son décodage linguistique dans le cerveau. C'est vrai pour les aspects segmentaux comme pour la prosodie. Regardons de plus près les trois types.

1. La *notation symbolique* se sert d'un petit inventaire de symboles, par exemple les niveaux de hauteur ou les tons. Elle ne retient de la prosodie que les éléments dits « pertinents » au niveau de l'organisation de l'énoncé, au niveau de la structuration informationnelle, etc. Typiquement, elle écarte les aspects rythmiques (tempo, accélérations, ralentissements, pauses, ...) et paralinguistiques (registre vocal, type de phonation, effort vocal), qui ont une fonction pragmatique ou phonostylistique. D'une manière générale, cette notation est tributaire du modèle adopté. On peut donc affirmer qu'elle est à la fois réductrice et partiale, de par sa nature. Cette notation est fournie manuellement par un transcripteur (qui maîtrise le modèle), ce qui comporte un risque de subjectivité évident. Il n'existe à ce jour aucun système automatique de transcription symbolique, malgré les nombreuses tentatives (par exemple, pour le français, MERTENS 1987, GEOFFROIS 1995, CAMPIONE, HIRST, VÉRONIS 2000).

2. Pour éviter la subjectivité et dans le but d'obtenir, de façon automatique, une représentation quantifiée, on peut bien sûr envisager une *analyse acoustique*. L'interprétation des tracés de fréquence fondamentale et d'intensité n'est cependant pas chose évidente : elle suppose leur alignement avec la transcription phonétique et présuppose de solides connaissances en phonétique acoustique afin d'identifier les phénomènes microprosodiques, les erreurs de détection de fondamental (saut d'octave), pour évaluer l'importance des intervalles mélodiques, et ainsi de suite. Il en résulte que ce type de représentation de la prosodie est peu lisible pour le non-spécialiste.

3. Rappelons-le : la représentation acoustique indique les paramètres acoustiques, mais elle ignore le traitement auditif et perceptif. Or, la suite de cet article démontrera précisément l'impact de la perception. Le fonctionnement du système auditif, appliqué aux propriétés acoustiques du signal de parole, entraîne le découpage du signal sonore en une suite de chaînons correspondant aux noyaux syllabiques. Par ailleurs, les *transcriptions auditives* de l'intonation proposées avant l'utilisation des mesures acoustiques en témoignent : elles représentent l'intonation comme une suite de contours syllabiques, le plus souvent plats dans le cas des syllabes atones (COUSTENOBLE & ARMSTRONG 1934, ZWANENBURG 1965). Une *représentation de l'intonation perçue*, telle que l'offre la transcription auditive manuelle, convient mieux aux besoins du linguiste que l'analyse acoustique, parce qu'elle se rapproche de l'image auditive à laquelle a accès l'auditeur.

Cependant, comme le rappellent CAMPIONE & VERONIS (2001 : 125), la transcription auditive manuelle « fait appel à une compétence phonétique très spécialisée peu courante parmi les 'linguistes de corpus'. De plus, la transcription prosodique est d'une nature éminemment subjective, qui réduit la fiabilité des données résultantes et impose des relectures par des annotateurs multiples accroissant encore le coût global

de la tâche ». Il est donc nécessaire d'automatiser la transcription prosodique, dans la mesure du possible, afin d'atteindre l'objectivité dans un laps de temps raisonnable.

Cette automatisation ne peut se faire au prix de l'information prosodique elle-même. La transcription obtenue ne peut pas être «large» au point d'amalgamer des contours intonatifs distincts et fonctionnels.

Précisons les objectifs d'un système de transcription prosodique. 1. Le but visé est une représentation *objective* et fiable de la prosodie, *facile à lire et à interpréter*. Cette transcription représentative de l'intonation perçue permettra de distinguer variations mélodiques audibles et inaudibles, au niveau des syllabes individuelles comme au niveau des séquences de syllabes. 2. En même temps, elle préservera l'évolution de la hauteur sur des fragments de parole plus longs, à un *niveau plus global*, permettant ainsi d'identifier les phénomènes de déclinaison, d'attaque, de registre et de changement de registre. 3. Tout ceci suppose en même temps que l'affichage de la hauteur soit *quantifié*, afin de permettre l'évaluation des intervalles mélodiques à chaque niveau (local ou global). 4. La transcription préservera l'*organisation temporelle*, afin de repérer et d'évaluer les pauses et hésitations, de déterminer le tempo (débit), d'étudier les aspects rythmiques, les accélérations et les ralentissements. 5. Vu la taille des corpus à analyser, une telle transcription devrait autant que possible être *automatique*. 6. Il importe aussi que la transcription soit *neutre*, c'est-à-dire indépendante de telle ou telle théorie de l'intonation. Cette neutralité va permettre son utilisation par des chercheurs d'horizons théoriques divers. 7. Afin de faciliter la consultation, la transcription comporte des *annotations phonétique et textuelle* alignées avec le signal. 8. Le caractère quantifié des dimensions hauteur et temps autorisera en outre des *manipulations*, par exemple en resynthèse (PSOLA) et en synthèse, permettant d'évaluer sa validité.

Dans la suite de cet article, nous présentons un système de transcription qui atteint en grande partie ces objectifs. Ce système fait intervenir une méthode de stylisation mise au point antérieurement (D'ALESSANDRO & MERTENS 1995). Sa particularité réside dans le fait qu'elle est basée sur une simulation de la perception de la hauteur et qu'elle prend comme unité de base le noyau syllabique. La transcription prosodique proposée sera appelée *prosogramme*, par analogie avec les termes oscillogramme et spectrogramme, qui représentent l'évolution de l'onde sonore et celle du spectre, dans le temps, respectivement.

Les figures ci-dessous offrent une première idée de la transcription proposée. Celle-ci se veut une estimation de la hauteur perçue par l'auditeur moyen. Plusieurs variantes ont été prévues. La transcription *simple* ne donne que la hauteur perçue (trait épais) ; la transcription *riche* montre en outre la fréquence fondamentale (trait fin en noir) et l'intensité (trait fin en gris). Ces informations peuvent être présentées sous deux formats : le format *compact*, prévu pour la transcription de corpus, et le format *large*, qui inclut une calibration des axes de temps (en s) et de fréquence (en demi-tons). En tout, on obtient donc quatre variantes.

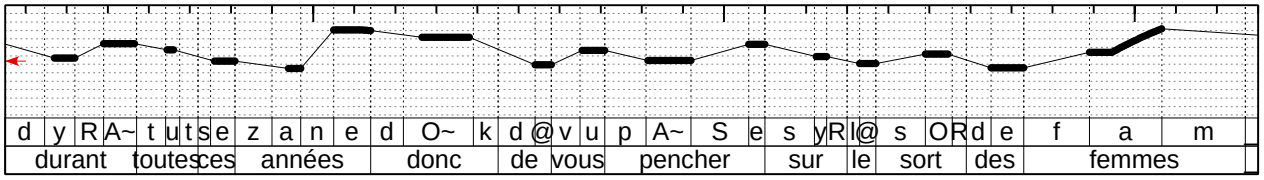


Figure 1. Prosogramme compact simple (seuil de glissando 0.32/T², cfr infra)

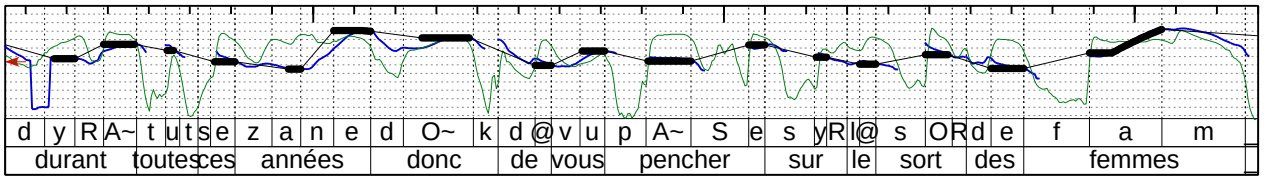


Figure 2. Prosogramme compact riche (seuil de glissando 0.32/T², cfr infra)

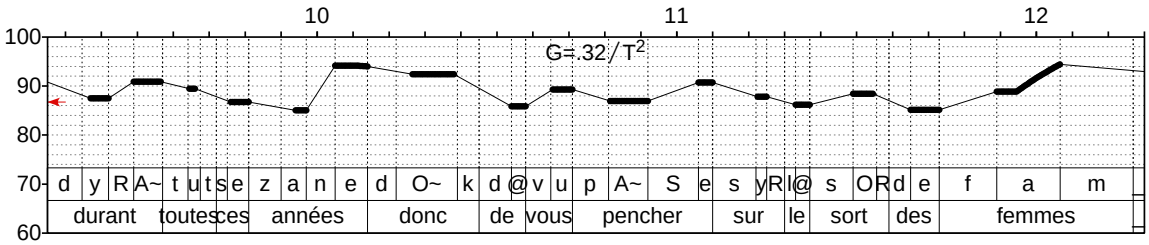


Figure 3. Prosogramme normal simple (seuil de glissando 0.32/T², cfr infra)

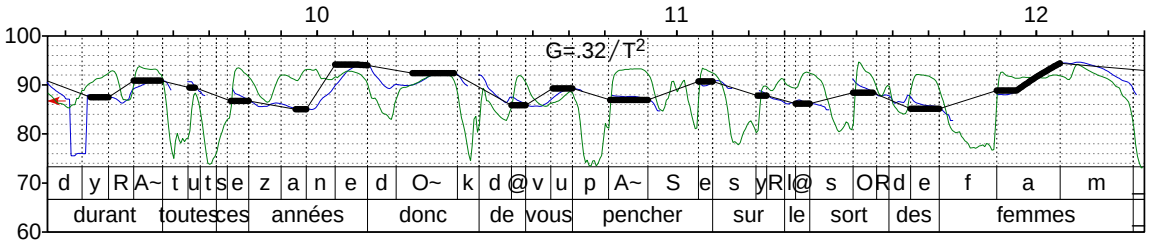


Figure 4. Prosogramme normal riche (seuil de glissando 0.32/T², cfr infra)

Au-dessus des annotations phonétique (au format SAMPA) et textuelle, on trouve un ensemble de lignes parallèles (en pointillé), distantes l'une de l'autre de 2 demi-tons (st, semitones). Cette portée musicale permet d'interpréter la hauteur des voyelles donnée par les traits épais, et d'évaluer les intervalles mélodiques soit entre syllabes, soit à l'intérieur des voyelles.

Par exemple, un intervalle (montant) de 4 st sépare la première syllabe du mot « pencher » de la deuxième. Comme les deux syllabes sont représentées par un trait plat, elles sont perçues sans variation de hauteur *interne*. Il en est autrement de la

syllabe « femmes », qui est représentée par une ligne inclinée ; on estime cette montée interne à 8 st.

La flèche sur le bord gauche indique la fréquence de 150 Hz (soit 86.75 st, relatif à 1Hz), et constitue une clef pour l'interprétation de la hauteur dans le prosogramme compact. La clef a été choisie à 150 Hz parce que cette valeur sera comprise dans le registre de la plupart des locuteurs, hommes ou femmes. L'échelle de fréquence en demi-tons prend ici comme référence la valeur de 1 Hz.

Les transcriptions riches donnent une vue plus détaillée, grâce à la courbe d'intensité. La courbe de fréquence fondamentale (affichée sur la même échelle en demi-tons) permet en outre d'expliquer et de valider la transcription obtenue.

Plusieurs prosogrammes seront commentés en détail dans les sections suivantes.

La suite de cet article précisera les considérations qui sous-tendent la transcription proposée, ses fondements et sa réalisation. La section 2 donne un aperçu de la perception de la hauteur dans la parole. La section 3 montre comment ces observations sont utilisées dans la stylisation automatique. Son application à la transcription prosodique fait l'objet de la section 4. Après les sections 5 (implémentation) et 6 (utilisation), la section 7 aborde la question de la validité (représentativité) et de l'évaluation du prosogramme. Celui-ci est confronté à la transcription manuelle du même passage. La méthode de la resynthèse de la parole avec la mélodie stylisée est utilisée ensuite pour valider le résultat par un test d'écoute informel. Nous terminons par la conclusion et les perspectives de recherche (§6).

2. LA PERCEPTION DE LA HAUTEUR

Dans la communication parlée, les contours intonatifs sont interprétés par un auditeur, et non pas par une machine. Le système auditif fonctionne autrement que les analyseurs spectraux ou que les détecteurs de mélodie utilisés en traitement du signal.

La psycho-acoustique étudie la relation entre telle ou telle propriété acoustique et son effet au niveau de la perception auditive. Dans le domaine des variations de hauteur dans la parole, plusieurs phénomènes perceptuels ont été identifiés. Nous nous limitons ici à une présentation sommaire de ces phénomènes ; ils sont discutés en détail dans D'ALESSANDRO & MERTENS (1995).

1. Pour être audible, une variation de fréquence fondamentale (F_0) doit présenter une ampleur minimale qui varie en fonction de la fréquence de départ et de la durée du stimulus (elle décroît avec la durée). Ce *seuil de glissando* a été évalué pour des variations de fréquence linéaires pour des sons purs, des sons de parole synthétiques, ou des sons de parole naturels resynthétisés (afin d'obtenir la variation linéaire). 'T HART (1976) propose une définition unifiée du seuil de glissando, où l'ampleur de la variation s'exprime comme un intervalle en demi-tons, ce qui permet d'éliminer le facteur fréquence de départ. Le seuil déterminé dans les expériences psycho-

acoustiques est de $G = 0.16/T^2$ (st/s), avec T la durée de la variation, cfr aussi les travaux récents de D'ALESSANDRO *et al*. (1998). (L'échelle en demi-tons correspond à l'échelle musicale : l'octave se divise en 12 intervalles égaux sur une échelle logarithmique.)

2. Bien sûr, les variations de fréquence dans la parole naturelle sont rarement linéaires. Comment sont perçues les variations comportant un changement de pente ? On peut reformuler la question comme suit : à partir de quel moment un changement de pente est-il audible ? D'ALESSANDRO & MERTENS (1995) proposent la notion de *seuil de glissando différentiel* (DG). Tout changement de pente est comparé au seuil DG et s'il est inférieur au seuil, le changement de pente n'est pas audible, et les deux parties concernées sont remplacées par une seule variation linéaire allant du début de la première à la fin de la deuxième. (Il existe peu de travaux sur ce seuil. La valeur utilisée ici est DG = g2 - g1 = 20 st/s, où g1 et g2 indiquent les pentes (en st/s) des deux parties de la variation, de part et d'autre du point de changement de pente.)

Il existe peu de travaux sur la perception des variations complexes : montant-descendant, montant-palier, ... (voir cependant les travaux de ROSSI 1978a, 1978c). On fera l'hypothèse que si chacune de ses parties est audible, la variation complexe sera perçue comme la séquence des variations constitutives simples.

3. Jusqu'ici il a été question seulement des variations mélodiques au cours de sons isolés (typiquement, des sons vocaliques). Or, dans la parole les sons s'enchaînent et cet enchaînement va de pair avec des variations d'intensité et de voisement, et avec des changements importants au niveau spectral. L'alternance entre voyelles et consonnes (ou groupes consonantiques) entraîne, *dans la plupart des cas*, un pic d'intensité et de sonorité pendant la voyelle, qui se caractérise par une stabilité relative du spectre. La voyelle constitue alors le noyau syllabique. En revanche, les consonnes situées entre ces voyelles coïncident avec des creux d'intensité et peuvent donner lieu à des changement spectraux relativement rapides et importants. Le contraste entre voyelles et consonnes est le plus prononcé pour les occlusives et fricatives, alors que les liquides, nasales et semi-voyelles se rapprochent des voyelles. Les changements acoustiques majeurs se situent donc aux frontières syllabiques. Par exemple, dans le prosogramme ci-dessus, les voyelles [u] et [e] dans « toutes ces » constituent des pics d'énergie par rapport aux consonnes avoisinantes. La différence d'intensité entre [t] et [u], dans « toutes ces années », est plus importante que celle entre [e] et [z]. Mais le [a] et le [m] de « femmes » ont une intensité comparable et le [m] présente son pic d'énergie propre.

Les travaux de HOUSE (1990) sur la perception des variations mélodiques dans la parole montrent qu'une même variation du fondamental sera perçue différemment selon sa place par rapport aux frontières syllabiques. Si elle apparaît au cours de la voyelle, la variation sera audible compte tenu du seuil de glissando. Si elle est située en partie pendant la transition à la frontière syllabique, seule la partie sur la voyelle sera bien intégrée auditivement. Tout semble indiquer que les changements

simultanés d'intensité, de spectre et de voisement entravent l'intégration perceptive des variations mélodiques. Le phénomène est d'autant plus prononcé que les changements acoustiques sont importants. Ceci donne lieu à la *segmentation du continuum mélodique* en chaînons correspondant aux noyaux syllabiques.

4. Dans la chaîne parlée, les sons et syllabes se suivent à toute allure et l'information mélodique doit être traitée en temps réel puisque d'autres sons arrivent déjà. La perception tonale est plus « performante » pour des voyelles présentées isolément que dans la parole continue à débit élevé : le seuil de glissando est plus élevé dans la parole continue. HOUSE (1995) montre en outre que les variations de fondamental sont mieux perçues quand elles sont suivies d'une *pause*. Autrement dit, la présence d'une pause après la variation abaisse le seuil de glissando.

3. LA STYLISATION DES VARIATIONS DE HAUTEUR ET LA PERCEPTION TONALE

Le terme de *stylisation* indique une forme simplifiée de la courbe de F0 qui est censée préserver les phénomènes fonctionnels ou audibles. L'idée semble remonter au travaux de J. 't Hart à l'IPO (Institut pour la Recherche sur la Perception, à Eindhoven, aux Pays-Bas). La « close copy stylization » est obtenue de façon interactive par la resynthèse du signal à partir de la courbe de F0 fournie par l'utilisateur. Cette courbe se présente comme une chaîne de segments de droite. L'utilisateur ajoute ou déplace des points jusqu'à ce qu'il soit impossible de distinguer la stylisation de l'original resynthétisé. Dans la « standardized pitch movement stylization », la courbe est constituée de l'enchaînement de mouvements standardisés tirés d'un inventaire prédéfini d'une dizaine de mouvements (toujours linéaires) ('T HART *et al*. 1990).

Les formes de stylisation abondent ; ce sont surtout les approches (semi-) automatiques qui retiendront l'attention ici. Dans les années 1980 et 1990, on a essayé d'obtenir de façon automatique une stylisation par mouvements standardisés ('T HART 1979a, SPAAI *et al*. 1993). Le système Momel (HIRST & ESPESSER 1993, CAMPIONE *et al*. 2000) modélise la courbe de F0 par une fonction de spline quadratique, comme une suite de segments de parabole. Le système proposé par RIETVELD (1984) utilise la régression linéaire pour déterminer les points d'inflexion de la courbe. Ces méthodes de stylisation reposent donc souvent sur les propriétés mathématiques ou statistiques de la courbe de F0.

D'ALESSANDRO & MERTENS (1995) proposent une approche de stylisation basée sur la *simulation de la perception tonale*. Nous donnons ici une description simplifiée de l'algorithme ; pour une présentation détaillée nous renvoyons le lecteur à l'ouvrage cité. Notons que les paramètres du modèle sont des seuils psycho-acoustiques.

1. *Segmentation* de la courbe. La courbe du fragment à analyser est d'abord subdivisée en segments temporaires de pente uniforme et relativement linéaire (le critère retenu est l'écart maximal entre les valeurs de F0 mesurées et la droite entre les valeurs au début et à la fin du segment). Chaque paire de segments temporaires contigus pour lesquels la différence de pente est inférieure au seuil de glissando différentiel sera remplacée par un seul segment.

2. *Stylisation*. Pour chaque segment retenu on évalue ensuite la variation mélodique. Si elle est inférieure au seuil de glissando, le segment est remplacé par une ligne horizontale à une fréquence égale au point d'arrivée du segment original. Dans le cas contraire, le segment sera remplacé par une droite reliant les valeurs initiale et finale du segment. Il en résulte que seules les variations audibles apparaissent comme des lignes inclinées dans la stylisation, alors que les changements de fréquence inaudibles donnent une ligne horizontale.

3. *Choix de l'unité de traitement*. Cette procédure peut être appliquée à toute portion voisée du signal de parole, quelle que soit sa longueur. Mais le choix de la portion analysée permet de tenir compte de l'effet de segmentation décrit plus haut.

Dans le modèle initial de 1995, la procédure était appliquée soit aux portions voisées de longueur maximale (portant éventuellement sur plusieurs syllabes), soit à la portion voisée de chaque syllabe. Cependant, les observations faites plus haut, à propos de l'effet de segmentation, suggèrent comme unité optimale la partie voisée du *noyau* syllabique, qui a effectivement été utilisée dans MERTENS & D'ALESSANDRO (1995). L'intérêt du noyau syllabique comme unité de base est double : il permet de localiser les perturbations microprosodiques à l'attaque de la syllabe et donc de les éliminer ; il permet un traitement plus adéquat des consonnes sonantes de la coda (liquides, nasales, semi-voyelles, fricatives sonores). Des procédures automatiques de segmentation en noyaux syllabiques ont été proposées (MERTENS 1987a, 1987b), mais elles reposent sur les propriétés acoustiques et ne sont pas assez robustes.

La méthode de stylisation décrite a été évaluée (D'ALESSANDRO & MERTENS 1995, MERTENS *et al.* 1997) dans des tests perceptifs où les sujets devaient discriminer deux stimuli synthétisés (par la technique PSOLA), l'un avec les valeurs originales de F0, l'autre avec la courbe stylisée, et ceci pour plusieurs valeurs des seuils G (0.16, 0.32, 0.64) et DG (20, 60). Cette méthode permettait de voir à partir de quelle valeur des seuils, la modification introduite par la stylisation devenait audible. (La même méthode permet d'établir le seuil de glissando en parole continue.) Comme le montre le prosogramme riche, la stylisation suit de près le F_0 : quand le trait épais de la stylisation couvre (et cache) le trait fin du F_0, la stylisation est identique au F_0 à 1 demi-ton près.

La stylisation s'écarte sur un point du modèle original décrit dans D'ALESSANDRO & MERTENS (1995). Dans celui-ci, une fonction était appliquée à la courbe de F0 avant la stylisation proprement dite. Sa fonction était de prendre en

compte le passé récent de la courbe de F0 pour la détermination de la hauteur perçue (WTA, *windowed time average pitch*). Ce traitement introduit cependant un lissage qui peut renforcer les phénomènes de microprosodie coarticulatoire et affecter de façon négative la stylisation finale.

4. APPLICATION DE LA STYLISATION À LA TRANSCRIPTION AUTOMATIQUE

La stylisation basée sur la perception tonale est mise en oeuvre pour la transcription de la prosodie, plus précisément comme représentation de la prosodie perçue.

En l'absence d'une segmentation automatique en noyaux syllabiques basée sur des critères perceptuels, on adoptera une solution pragmatique qui consiste à modéliser la courbe mélodique des *voyelles* seulement. L'information nécessaire provient de l'alignement phonétique, donnant, pour chaque son dans le signal, le symbole phonétique et les instants temporels du début et de la fin.

Pour chaque voyelle on détermine ensuite le *noyau vocalique*. Celui-ci est défini comme la partie *voisée* autour du pic d'intensité, *délimitée* à gauche et à droite par les points situés à -3 dB et -9 dB du maximum, respectivement. La valeur pour la frontière gauche permet d'éliminer en partie les perturbations microprosodiques à l'attaque syllabique (elles peuvent être considérables dans le cas des obstruantes sourdes) et d'éviter les phénomènes microprosodiques sur les consonnes voisées à la jonction entre deux syllabes ; la valeur pour la frontière droite permet de préserver les variations de fréquence tardives (dans le cas des voyelles accentuées).

Il est clair que le résultat obtenu dépend en grande partie de la qualité de l'alignement phonétique utilisé.

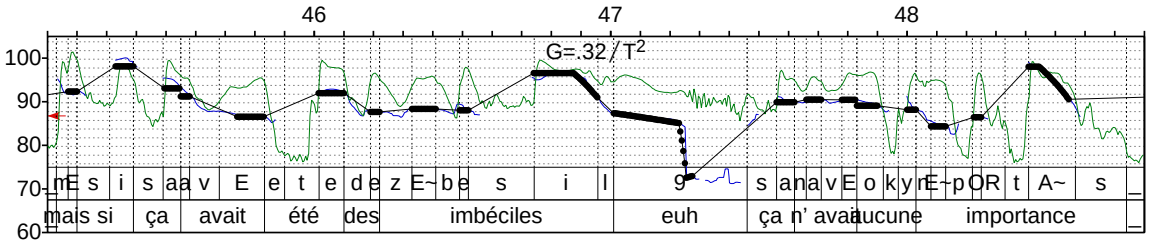


Figure 5. Illustration de la délimitation du noyau vocalique (voir texte)

La figure 5 permet d'illustrer l'importance de la délimitation du noyau syllabique. La voyelle finale du mot « importance » présente une chute à grand intervalle (-10 st) depuis le niveau haut (ton HB). Le choix de la frontière droite à -9 dB a permis de préserver cette chute dans sa presque totalité et d'identifier ainsi le ton HB. Le mot « imbéciles » présente le même type de contour, qui se réalise en

partie sur la consonne finale. Le « euh » d'hésitation suivant prolonge le niveau bas qui descend progressivement jusqu'à ce que la vibration glottale devienne irrégulière (fait visible sur la courbe d'intensité), provoquant un saut vers le bas dans le tracé de F0 et dans la stylisation résultante. La délimitation du noyau vocalique a permis d'éliminer la presque totalité de la phase irrégulière.

Le seuil de glissando est un paramètre du modèle. Comme il a été mentionné plus haut, ce seuil dépend de la nature du signal : son isolé ou parole continue. Comment déterminer le seuil de glissando adéquat ?

La stylisation obtenue a été comparée systématiquement avec la transcription manuelle d'un corpus test (le corpus Fayard-Groult, utilisé au colloque de Genève, en septembre 2002), effectuée préalablement par deux annotateurs expérimentés. Examinons d'abord le traitement des variations locales (des syllabes ou des groupes). Avec un seuil de glissando de $G = 0.16/T^2$, la stylisation retient plus de variations intrasyllabiques que la transcription manuelle. Autrement dit, la stylisation avec le seuil standard, observé dans les expériences psycho-acoustiques portant sur des stimuli présentés isolément, surestime les capacités de l'auditeur moyen. Pour un seuil de $G = 0.32/T^2$, soit deux fois le seuil standard, la stylisation est très proche de la notation manuelle, pour ce qui est du choix entre glissando et ton statique. Ces données confirment l'élévation du seuil dans la parole continue. Quant aux variations mélodiques globales, qui s'étendent sur plusieurs secondes, la transcription semi-automatique s'avère plus précise que la transcription manuelle. En effet, le seuil de glissando n'affecte pas la perception de variations de hauteur portant sur plusieurs syllabes.

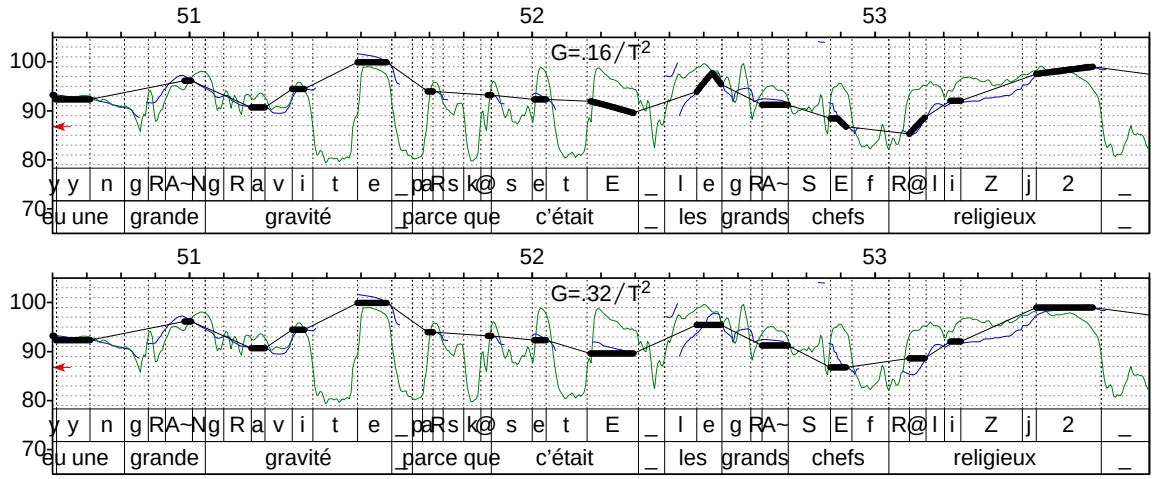


Figure 6 . Transcription prosodique pour deux valeurs différentes du seuil de glissando G.

La figure 6 illustre l'effet du seuil de glissando G utilisé pour la stylisation. La partie supérieure, qui utilise le seuil bas, comporte davantage de variations intrasyllabiques jugées audibles (par exemple, les syllabes « tait » et « gieux ») que la

partie inférieure où est employé le seuil élevé. Dans le cas de la syllabe « chefs », la variation est clairement due à un phénomène microprosodique. La syllabe « les » porte un accent initial (ou accent d'insistance) qui se manifeste par la tenue particulièrement longue de la consonne initiale et, du moins si celle-ci est voisée, par une variation de fréquence F0 à la fois rapide et importante.

5. IMPLÉMENTATION

Dans sa forme actuelle le système de transcription fait intervenir plusieurs logiciels. Le logiciel Praat est utilisé pour le calcul des paramètres acoustiques (fréquence fondamentale, intensité, décision de voisement) et pour l'affichage des résultats (stylisation, annotation phonétique et textuelle, fréquence fondamentale, intensité). Tous ces aspects, ainsi que l'interface utilisateur, sont décrits dans un script principal en langage Praat. La segmentation en noyaux vocaliques est effectuée par un script en Perl, qui prend comme entrées l'annotation phonétique et les paramètres d'intensité et de voisement. La stylisation tonale proprement dite est effectuée par un logiciel indépendant, écrit en langage C. La communication entre ces composantes s'effectue par fichiers intermédiaires ; les nombreuses conversions de format qu'elle suppose sont réalisées en Perl. Ces conversions sont lancées à partir du script principal.

À court terme les étapes de segmentation, de stylisation et de conversion de format seront remplacées par des scripts en Praat. L'ensemble de la procédure sera ainsi contenue dans un seul script, afin de faciliter l'installation et d'éliminer les limitations de portabilité.

Afin d'obtenir une précision maximale, les paramètres de F0, d'intensité et la décision de voisement sont calculés toutes les 5 ms ("frame rate" = 200 Hz).

Le calcul des paramètres demande environ 5 fois le temps du signal sur une machine à 450 MHz. Le calcul du prosogramme proprement dit se fait en temps réel.

6. UTILISATION

L'outil de transcription suppose deux fichiers d'entrée : le fichier son et l'annotation phonétique correspondante (au format TextGrid de Praat). Le nom des deux fichiers sera identique à l'exception de l'extension.

L'outil permet d'obtenir la transcription d'un extrait au choix ou d'un corpus entier. Dans le premier cas, on fournit le nom du fichier son et les temps du début et de la fin de l'extrait. Dans le mode de visualisation compact, une page A4 peut contenir 9 à 10 prosogrammes, soit environ 30 s de transcription.

L'interface utilisateur permet d'ajuster les paramètres suivants.

1. Le répertoire où se trouvent les fichiers à analyser (son et annotation phonétique).
2. Le nom du (des) fichier(s) son à analyser. L'utilisation de wildcard permet de sélectionner un ensemble des fichiers. Pour le traitement d'un corpus, les fichiers seront traités dans l'ordre, par exemple : A001.wav, A002.wav, et ainsi de suite.
3. Les temps au début et à la fin de l'extrait à transcrire. Les valeurs par défaut sont le début et la fin du signal de parole, respectivement.
4. La durée de la fenêtre de transcription. Elle est de 3 s par défaut.
5. Le mode d'affichage: simple ou riche, compact ou normal.
6. La plage de fréquences à utiliser pour l'axe vertical. Celle-ci peut être spécifiée en Hertz ou en demi-tons (relatif à 1 Hz).
7. Le répertoire de sortie, où seront sauvegardés les fichiers de prosogrammes, qui seront numérotés automatiquement.
8. Le préfixe à utiliser pour les fichiers de transcription.

L'annotation phonétique requise comme point de départ peut être obtenue de plusieurs façons. On peut la constituer manuellement, à l'aide d'un outil interactif comme Praat (TextGrid editor). Certains laboratoires de recherche disposent d'outils d'alignement automatiques, basés sur la reconnaissance automatique de la parole ou sur l'alignement du signal original avec le signal synthétisé à partir de la transcription textuelle ou phonétique qui sera alors fournie également en entrée.

Les fichiers de transcription (générés par Praat) au format graphique EPS (Encapsulated Postscript) peuvent être visionnés et imprimés à l'aide du logiciel Ghostview (un logiciel en accès libre) ou peuvent être imprimés directement sur une imprimante Postscript. Le logiciel Ghostview permet aussi de convertir les fichiers EPS vers d'autres formats graphiques (dont PDF, PNG, PS). Certains logiciels de traitement de texte permettent d'incorporer les illustrations au format EPS dans des documents. Le format EPS donne alors la qualité optimale.

7. DISCUSSION

Afin de valider le système, plusieurs corpus de parole ont été analysés, pour lesquels des transcriptions manuelles avaient été réalisées préalablement par des annotateurs expérimentés. Il s'agit du corpus B. Groult (émission « La ligne de cœur » de Roselyne Fayard, 13/06/1996, Radio suisse romande 1), utilisé au colloque Prosodie à Genève de septembre 2002, et des corpus R. Barthes (émission « Radioscopie » de J. Chancel, de 17/02/1975, environ 11 min.) et F. Giroud (« Radioscopie » de

15/09/1977, environ 9 min.), pour lesquels MERTENS (1987a) fournit une transcription manuelle.

La prosodie expressive de Benoîte Groult se caractérise par l'ampleur des intervalles mélodiques (donnant un registre large), par l'utilisation fréquente du niveau suraigu (plafond de la tessiture du locuteur), par l'exploitation de la phonation (qualité vocale) et de la variation du débit. Les voix de J. Chancel, R. Barthes, F. Giroud et R. Fayard présentent un registre modal.

La confrontation des transcriptions automatique et manuelle permet d'étudier leur degré de correspondance et donc de voir dans quelle mesure l'une est représentative de l'autre et dans quelle mesure le prosogramme reproduit l'image auditive. Regardons l'extrait ci-dessous, tiré du corpus Barthes ; les prosogrammes sont calculés pour le seuil $G = 0.32/T^2$.

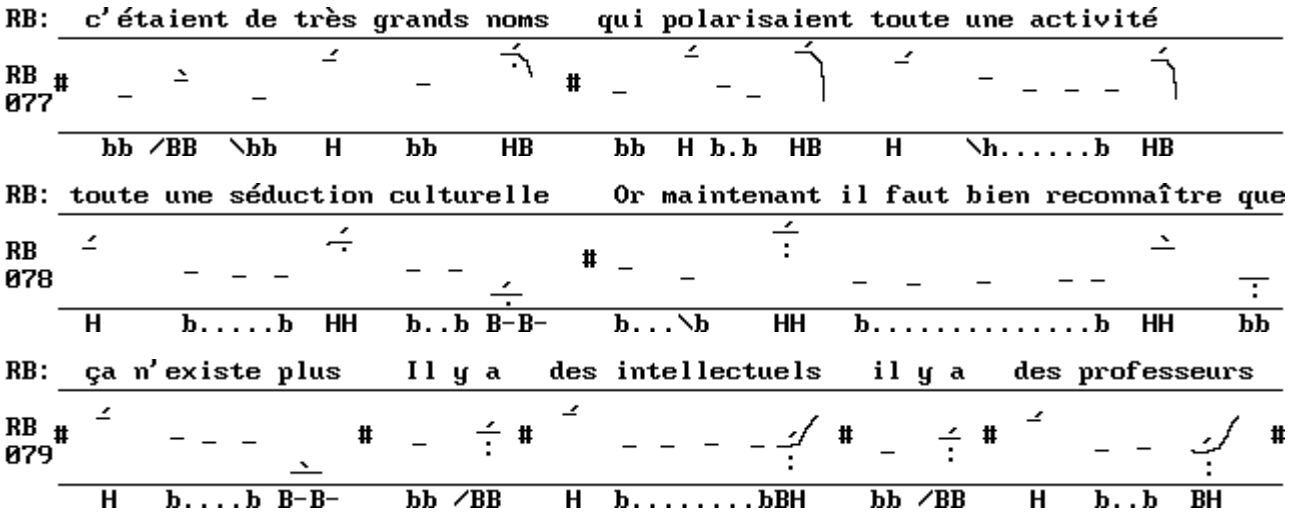
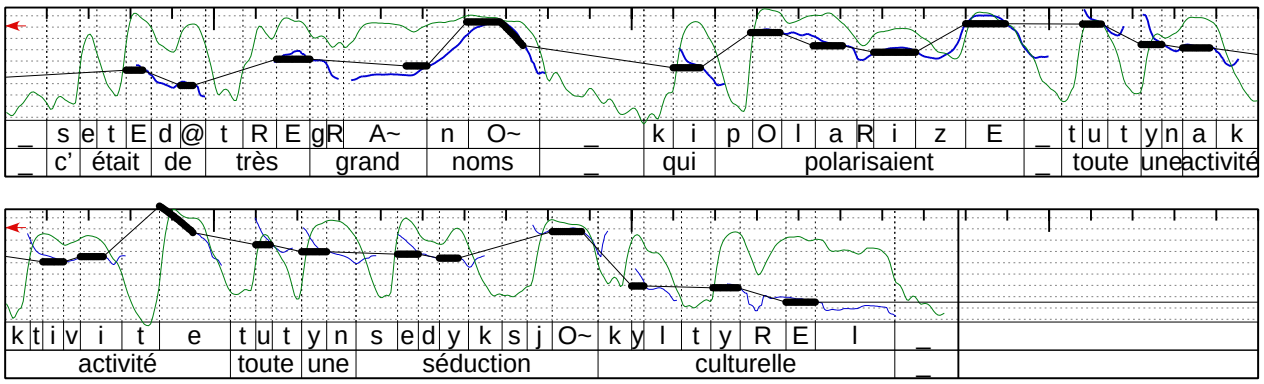


Figure 7. Transcription manuelle tirée du corpus Barthes (MERTENS 1987a)



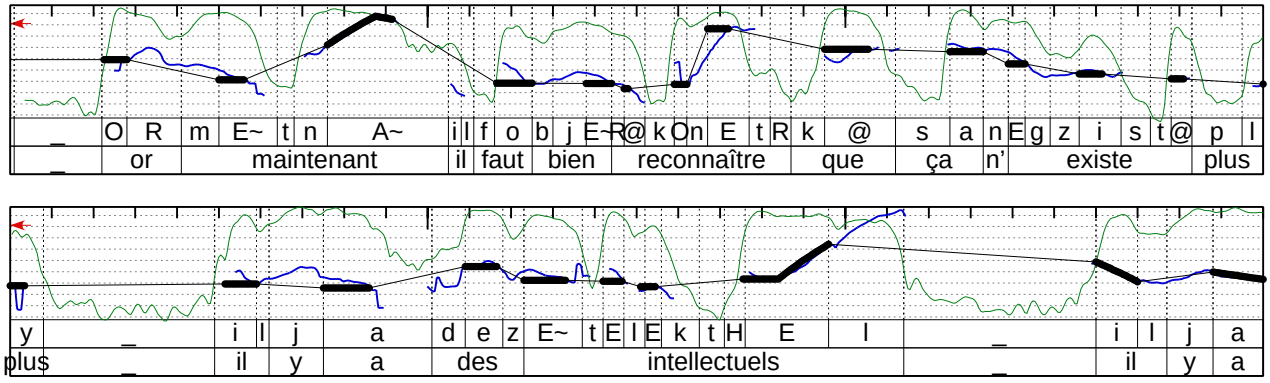


Figure 8. Prosogrammes de l'extrait de la figure 7.

Réalisées à quinze ans d'intervalle, les deux versions sont assez proches. Cela n'a rien d'étonnant, puisque les principes de transcription (la syllabe comme unité de base, la perception comme critère principal) sont identiques. Les deux s'accordent sur le caractère statique ou dynamique des syllabes, sur la direction et l'ampleur des glissandos. (Dans « polarisaient », le prosogramme rate la chute finale.) Les écarts sont plus importants pour les intervalles intersyllabiques.

Cependant les prosogrammes présentent plusieurs avantages majeurs. La nature quantifiée du prosogramme permet une évaluation directe des intervalles mélodiques, et des propriétés temporelles (débit, rythme, pauses). Le prosogramme constitue une procédure objective. Il permet un gain de temps énorme.

La méthode de stylisation n'entraîne aucune perte d'information. Le français connaît plusieurs contours de groupe intonatif qui se distinguent par l'endroit où se situe la montée et/ou par la présence éventuelle d'une descente après la montée. La notation proposée par MERTENS (1990) oppose ainsi les contours « b..b HH », « b..b H/H », « b..b BH », « b..b HB », « b..h BB » (le dernier présente une montée sur la syllabe pénultième atone), et ainsi de suite. Alors que bon nombre de « théories » de l'intonation du français ne les distinguent pas (ou pas tous), ces contours peuvent tous être distingués sur le prosogramme, comme le montrent les illustrations suivantes.

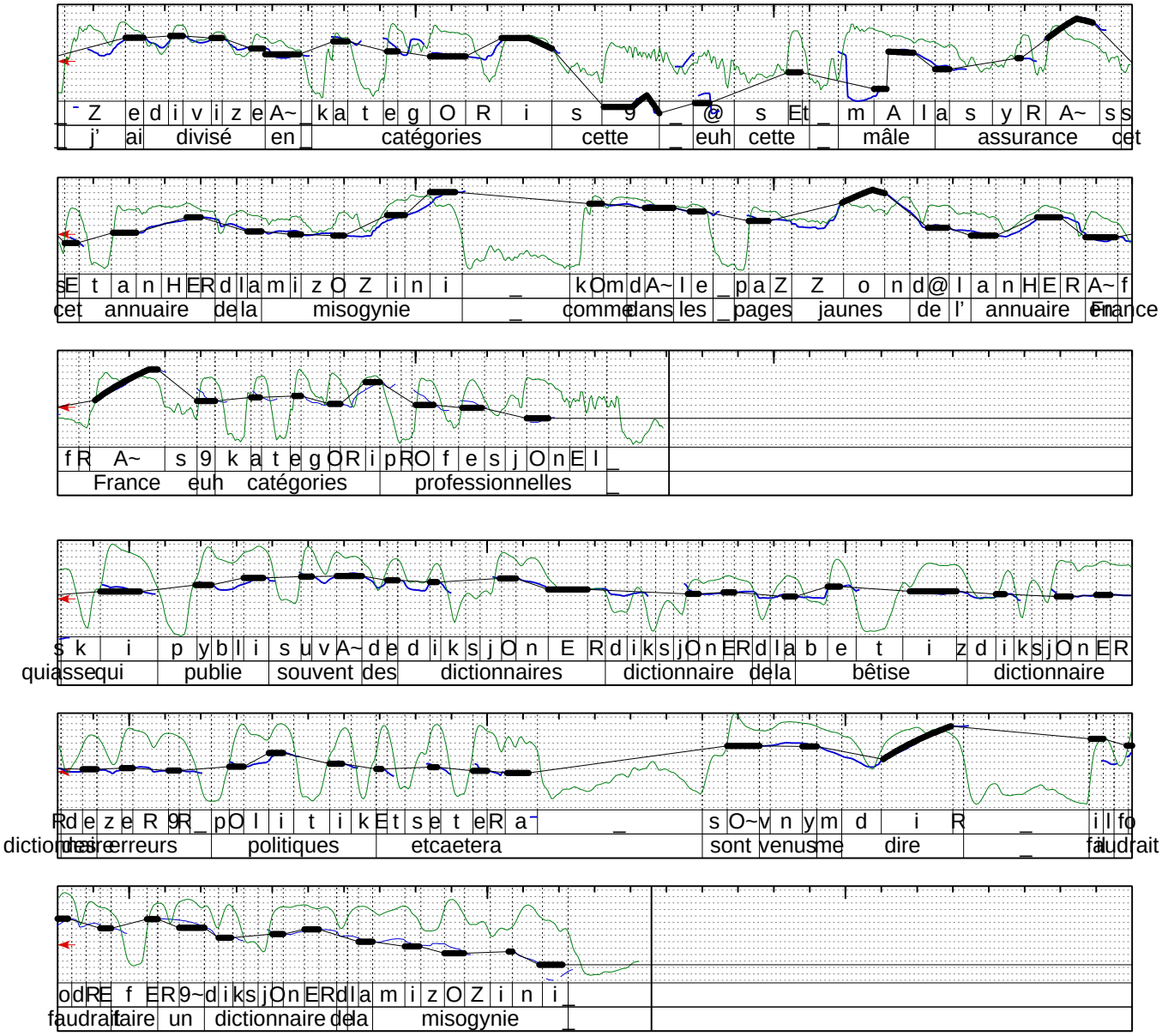


Figure 9. Prosogrammes de l'extrait Fayard-Groult

Afin de valider la stylisation, celle-ci a été utilisée pour resynthétiser le signal de parole. On crée un signal de parole qui a toutes les propriétés du signal original, sauf la fréquence fondamentale, pour laquelle on reprend la stylisation calculée. Pour les parties entre les noyaux vocaliques (il s'agit des parties non voisées et des consonnes), non traitées par la stylisation, la fréquence fondamentale utilisée correspond à l'interpolation linéaire entre les valeurs aux extrémités des noyaux avoisinants. La méthode utilisée, PSOLA, préserve l'organisation temporelle (durée des segments). Dans la plupart des cas, le signal resynthétisé peut difficilement être distingué du signal original. Des fragments resynthétisés figurent parmi les documents sonores qui accompagnent cet article.

8. CONCLUSION ET PERSPECTIVES

CAMPIONE & VERONIS (2001 : 123) résument clairement l'enjeu de la transcription prosodique automatique : « La transcription manuelle de la prosodie est une tâche extrêmement coûteuse en temps, qui requiert des annotateurs très spécialisés, et qui est sujette à de multiples erreurs et une grande part de subjectivité. Une annotation complète n'est pas envisageable dans l'état actuel de la technologie [...]. » Les outils de transcription permettent « une réduction substantielle du temps d'intervention manuelle, et améliorent l'objectivité et la cohérence du résultat. De plus, les étapes manuelles nécessaires ne demandent pas une expertise phonétique poussée et peuvent être menées à bien par des étudiants et des «linguistes de corpus» ».

Une approche de transcription prosodique semi-automatique a été présentée dans cet article. Ses particularités sont les suivantes. Elle se présente comme une stylisation de la courbe de F_0, pour les noyaux vocaliques, qui vise à reconstituer le contour mélodique perçu, en se basant sur un modèle psycho-acoustique de la perception tonale. La transcription prosodique, qui préserve la structure temporelle du signal acoustique, inclut des annotations textuelle et phonétique, la dernière étant utilisée pour l'identification des noyaux vocaliques. Les variations mélodiques apparaissent sur une échelle de hauteur en demi-tons ; le choix de l'échelle musicale répond à l'objectif de lisibilité. Grâce à l'alignement temporel, la transcription permet de déterminer la durée des sons et des syllabes, d'identifier les pause et de mesurer leur durée, et enfin d'étudier le débit et le rythme.

Par rapport aux approches qui visent une transcription symbolique et qui ne retiennent qu'un petit inventaire de symboles, la stylisation tonale est plus détaillée et évite toute prise de position sur la nature et l'inventaire des unités abstraites (contour, ton, groupe, etc.).

Deux formats de transcription ont été élaborés. Le format concis ne retient que la stylisation mélodique alignée avec les annotations phonétique et textuelle. Le format riche prévoit en outre le tracé de F0 (converti vers l'échelle musicale) et la courbe d'intensité. Il permet de valider la stylisation et d'étudier le rôle de l'intensité. Ces deux formats peuvent être présentés sous une forme compacte, en vue de la transcription prosodique de corpus.

L'outil a été utilisé pour transcrire trois corpus, faisant intervenir cinq locuteurs (deux hommes, trois femmes). Les résultats montrent la robustesse de la transcription, sa similarité avec la transcription manuelle, et la similarité des contours avec ceux utilisés en synthèse de la parole (MERTENS *et al*. 2001). La stylisation obtenue a été validée de façon informelle grâce à la resynthèse.

Plusieurs améliorations et extensions sont envisagées, comme l'affichage, sous forme graphique, du débit mesuré, ou l'ajustement automatique de la plage de hauteur affichée en fonction de la distribution des valeurs de F_0 mesurées dans le signal ou dans l'ensemble du corpus. À terme, le noyau vocalique devrait être remplacé par le

noyau syllabique comme unité de base pour la stylisation. Le noyau syllabique pourrait être défini à partir des propriétés acoustiques du signal ou à partir d'une syllabification de l'annotation phonétique éventuellement en combinaison avec un traitement automatique des informations textuelles.

Piet MERTENS
Département de Linguistique
K.U.Leuven
Blijde-Inkomststraat 21
3000 Leuven, Belgique
Piet.Mertens@arts.kuleuven.ac.be

RÉFÉRENCES BIBLIOGRAPHIQUES

ALESSANDRO, Ch. d', S. ROSSET & J.-P. ROSSI. 1998. « The pitch of short-duration fundamental frequency glissandos », *J. Acoust. Soc. Am.* 104/4, 2339-2348.

ALESSANDRO, Ch. d', P. MERTENS. 1995. « Automatic pitch contour stylization using a model of tonal perception », *Computer Speech and Language* 9/3, 257-288.

CAMPIONE, E., D. HIRST & J. VÉRONIS. 2000. « Stylisation and symbolic coding of F0: comparison of five models », in BOTINIS, A. (ed.), *Intonation : Analysis, Modelling and Technology*, Dordrecht : Kluwer Academic Publishing, 185-208.

CAMPIONE, E. & J. VÉRONIS. 2001. « Étiquetage prosodique semi-automatique des corpus oraux », *Actes TALN 2001*, 123-132.

COUSTENOBLE, H.N. & L.E. ARMSTRONG. 1934. *Studies in French intonation*, Cambridge : Heffer.

GEOFFROIS, E. 1995. *Extraction robuste de paramètres prosodiques pour la reconnaissance de la parole*, Thèse de doctorat, Université Paris XI Orsay, 20 décembre 1995.

't HART, J. 1974. « Discriminability of the size of pitch movements in speech », *I.P.O. Annual Progress Report* 9, 56-63.

't HART, J. 1976. « Psychoacoustic backgrounds of pitch contour stylization », *I.P.O. Annual Progress Report* 11, 11-19.

't HART, J. 1979. « Explorations in automatic stylization of F0 curves », *I.P.O. Annual Progress Report* 14, 61-65.

't HART, J., R. COLLIER & A. COHEN. 1990. *A perceptual study of intonation*, Cambridge : Cambridge Univ. Press.

HIRST, D. & R. ESPESSER. 1993. « Automatic Modelling of Fundamental Frequency Using a Quadratic Spline Function », *Travaux de l'Institut de Phonétique d'Aix-en-Provence* 15, 75-85.

HOUSE, D. 1990. *Tonal Perception in Speech*. Lund : Lund University Press.

HOUSE, D. 1995. « The influence of silence on perceiving the preceding tonal contour », *Proc. Int. Congr. Phonetic Sciences* 13, vol. 1, 122-125.

MERTENS, P. 1987a. *L'intonation du français. De la description linguistique à la reconnaissance automatique*. Thèse de doctorat, Université de Leuven.

MERTENS, P. 1987b. « Automatic segmentation of speech into syllables », *Proceedings of the European Conference on Speech Technology*, vol. 2, 9-12.

MERTENS, P. 1989. « Automatic recognition of intonation in French and Dutch », *Eurospeech* 89, vol. 1, 46-50.

MERTENS, P. 1990. « Chap. IV. L'intonation », in C. BLANCHE-BENVENISTE, M. BILGER, Ch. ROUGET & K. VAN DEN EYNDE. *Le français parlé : Études grammaticales*, Paris: Éd. du CNRS, 159-176.

MERTENS, P. & Ch. d'ALESSANDRO. 1995. « Pitch contour stylization using a tonal perception model », *Proc. Int. Congr. Phonetic Sciences* 13, 4, 228-231.

MERTENS, P., F. BEAUGENDRE & Ch. d'ALESSANDRO. 1997. « Comparing approaches to pitch contour stylization for speech synthesis », in J.P.H. VAN SANTEN, R.W. SPROAT, J.P. OLIVE & J. HIRSCHBERG (eds), *Progress in Speech Synthesis*. N.Y. : Springer Verlag, 347-363.

MERTENS, P., J.-P. GOLDMAN, É. WEHRLI & A. GAUDINAT. 2001. « La synthèse de l'intonation à partir de structures syntaxiques riches », *Traitement Automatique des Langues* 42/1, 145-192.

RIETVELD, A.C.M. 1984. *Syllaben, klemtonen en de automatische detectie van beklemtoonde syllaben in het Nederlands*. Thèse de doctorat, Université de Nijmegen.

ROSSI, M. 1971. « Le seuil de glissando ou seuil de perception des variations tonales pour la parole », *Phonetica* 23, 1-33.

ROSSI, M. 1978a. « La perception des glissandos descendants dans les contours prosodiques », *Phonetica* 35/1, 11-40.

ROSSI, M. 1978b. « The perception of non-repetitive intensity glides on vowels », *Journal of Phonetics* 6/1, 9-18.

ROSSI, M. 1978c. « Interactions of intensity glides and frequency glissandos », *Language & Speech* 21, 384-396

SPAAI, G.W.G., A. STORM, A.S. DERKSEN, D.J. HERMES & E.F. GIGI. 1993. « An Intonation Meter for teaching intonation to profoundly deaf persons », *IPO Manuscript* no. 968.

ZWANENBURG, W. 1965. *Recherches sur la prosodie de la phrase française*. Leiden : Universitaire Pers.