

---

# Un outil pour la transcription de la prosodie dans les corpus oraux

**Piet Mertens**

*Université de Leuven, Département de Linguistique  
21, Blijde-Inkomststraat 21  
3000 Leuven (Belgique)  
Piet.Mertens@arts.kuleuven.ac.be*

---

*RÉSUMÉ. On présente un outil d'aide à la transcription de la prosodie dans les corpus oraux. Il vise une transcription prosodique lisible, objective, quantifiée, semi-automatique, perceptuellement motivée, indépendante de telle ou telle théorie de l'intonation. Il s'agit d'une stylisation de la courbe de  $F_0$  basée sur un modèle de la perception tonale existant, qui est appliquée ici aux voyelles. Les variations infraliminaires apparaissent comme des traits plats, les glissandos comme des courbes. Pour faciliter l'estimation des intervalles mélodiques, le tout est affiché sur une échelle musicale en demi-tons, similaire à une partition musicale. La transcription est synchronisée avec la segmentation phonétique ; d'autres couches d'annotation peuvent être ajoutées. Le résultat a été comparé aux transcriptions manuelles d'experts ; elle peut être validée par resynthèse du signal sonore à partir de la stylisation.*

*ABSTRACT. We present a tool for prosody transcription of speech corpora. It aims at a transcription of prosody that is readable, objective, quantified, semi-automatic, perceptually motivated, and theory-independent. It uses a pitch contour stylization based on a tonal perception model proposed earlier. The stylization is applied to vowels in the signal. Subliminal variations appear as level lines and glissandos as curves. To facilitate pitch interval estimation, the result is plotted on a musical scale in semitones, similar to a musical score. The transcription is synchronized with the phonetic segmentation; additional annotation tiers may be displayed. The result was confronted with manual transcriptions of expert transcribers. As the stylization is available, it can be validated through resynthesis.*

*MOTS-CLÉS : Prosodie, Annotation, Corpus oraux, Stylisation, Perception tonale*

*KEYWORDS: Prosody, Annotation, Speech corpora, Stylization, Tonal perception*

---

## 1. Introduction

L'utilisation à grande échelle de corpus de langue, ces dernières décennies, a permis à la linguistique non seulement de réaliser des progrès importants, mais aussi de défricher des domaines de recherche nouveaux. Le traitement automatique des corpus n'a fait qu'accélérer cette évolution. Les corpus oraux mettent en évidence l'usage réel de la langue, plus diversifié que la norme. S'il existe un domaine linguistique où l'utilisation de corpus peut apporter beaucoup, c'est sans doute celui de l'étude de la prosodie, plus particulièrement celui de l'intonation. Le présent article présente un outil qui permet de réaliser des transcriptions semi-automatiques de la prosodie d'énoncés et qui a été conçu spécialement pour le traitement de corpus oraux.

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. 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. Toutefois, comme la transcription prosodique présuppose une compétence très spécialisée et représentait un investissement en temps prohibitif, l'intégration de l'intonation était toujours remise aux calendes grecques. L'utilisation d'un outil de transcription semi-automatique pourrait enfin donner naissance à des corpus de langue parlée explicitant l'intonation et faire apprécier à sa juste valeur son rôle dans la communication parlée.

### 1.1. *Quelle transcription de la prosodie ?*

Quel genre de transcription adopter pour annoter un corpus ? Grosso modo on peut distinguer trois types majeurs de représentations de la prosodie : l'analyse acoustique (paramètres de fréquence fondamentale, d'intensité et le spectre), la notation auditive et enfin la notation symbolique de type phonologique ou tonologique. Ces trois types correspondent à des étapes (de représentation et de traitement) dans la communication parlée. En effet, le signal sonore, de nature acoustique, passe d'abord (et inévitablement) par un traitement auditif et perceptif, 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 d'analyse adopté. On peut donc affirmer qu'elle est à la fois réductrice et partielle, de par sa nature.

Cette notation est fournie manuellement par un transcripateur (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 *et al.*, 2000 ; autres travaux : Beaugendre *et al.*, 1996 ; Taylor, 1993).

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. La représentation acoustique indique les paramètres acoustiques, mais elle ignore le traitement auditif et perceptif. Or, la suite de cet article rappellera 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'introduction des appareils de mesure spécialisés en témoignent : elles représentent l'intonation comme une suite de contours syllabiques, le plus souvent plats dans le cas des syllabes atones (Coustenoble *et al.*, 1934, pour le français ; les différents auteurs de l'école britannique, pour l'anglais). 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 *et al.*, 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 fonctionnels distincts.

### **1.2. Besoins et objectifs**

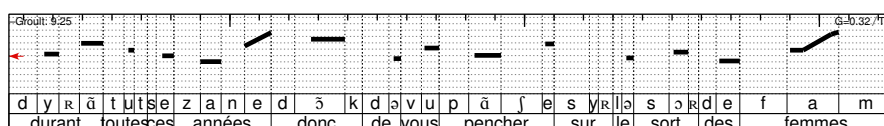
Précisons les objectifs d'un système de transcription prosodique. 1. Le but visé est une représentation *objective* et fiable de la prosodie, *facile à lire et à interpréter*. Cette transcription représentative de l'intonation perçue permettra de distinguer les variations de fréquence fondamentale audibles 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 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 *et al.*, 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* (et l'outil le *prosographe*), par analogie avec les termes oscillogramme et spectrogramme, qui représentent l'évolution de l'onde sonore et celle du spectre, dans le temps, respectivement.

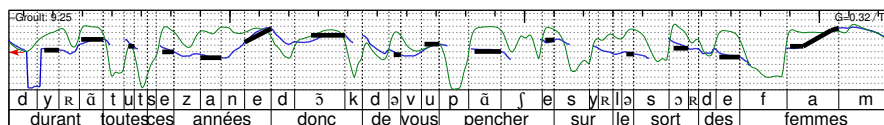
### **1.3. Présentation de la transcription**

Les figures ci-dessous offrent une première idée de la transcription proposée. Plusieurs variantes ont été prévues. La transcription *simple* ne donne que la hauteur stylisée (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* et le format *large*. Ce dernier inclut une calibration des axes de temps (en s) et de fréquence (en demitons), ce qui semble souhaitable pour les illustrations de publications, fournissant la transcription de passages relativement courts. Le format compact est prévu pour la transcription d'un corpus dans son ensemble. Grâce à ses dimensions réduites, il

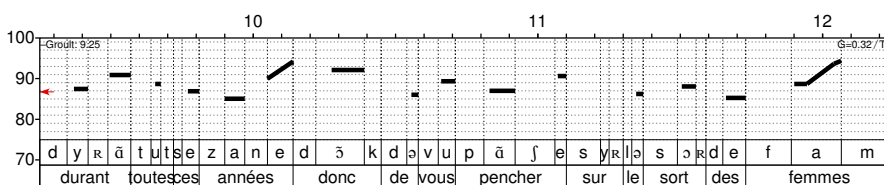
permet d'afficher un maximum de temps de parole : une page A4 peut tenir 9 bandes (ou prosogrammes) à deux couches d'annotation (la transcription phonétique et le texte), soit 27 s de parole (pour la durée par défaut de 3 s par bande), là où le format large ne permettrait que 6 bandes (soit 18 s). En tout, on obtient donc quatre variantes.



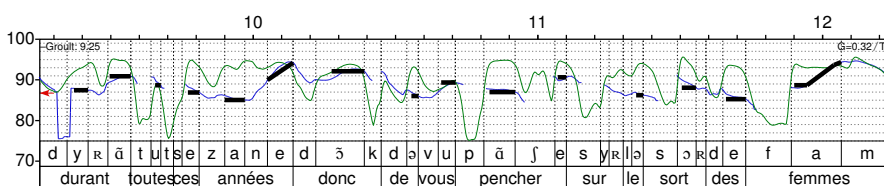
**Figure 1.** Prosogramme compact simple (seuil de glissando  $G = 0.32/T^2$ )



**Figure 2.** Prosogramme compact riche (seuil de glissando  $G = 0.32/T^2$ )



**Figure 3.** Prosogramme large simple (seuil de glissando  $G = 0.32/T^2$ )



**Figure 4.** Prosogramme large riche (seuil de glissando  $G = 0.32/T^2$ )

Au-dessus des annotations phonétique 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 7 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 hauteur en demi-tons prend ici comme référence la valeur de 1 Hz :  $f$  (en ST) =  $12 * \log_2(f/f_{ref})$ .

Les transcriptions riches donnent une vue plus détaillée, grâce aux courbes d'intensité et de fréquence fondamentale (affichée également en demi-tons), qui permettent 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. La section 5 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. La section 6 commente les fonctionnalités de l'outil prévues afin d'obtenir la transcription prosodique de corpus oraux entiers avec le minimum de manipulations. Nous terminons par la conclusion et par les perspectives de recherche (§7).

## 2. La perception de la hauteur

Dans la communication parlée, les contours intonatifs sont interprétés par un auditeur humain, 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 *et al.*, 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) (Rossi, 1971 ; 't Hart, 1974). ('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 psychoacoustiques est de  $G = 0.16/T^2$  (ST/s), avec T la durée de la variation, cf. 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 :  $12 \cdot \log_2(f_2/f_1)$ . Voir aussi Hermes *et al.*, 1991.)

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  $DG = g_2 - g_1 = 20$  ST/s, où  $g_1$  et  $g_2$  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 changements 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 à l'attaque syllabique et à la coda.

Par exemple, dans le prosogramme riche ci-dessus (figure 4), 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**

#### **3.1. Approches initiales**

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 aux travaux de J. 't Hart. La “close copy stylization” avancée par cet auteur 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 suite 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. Le système proposé par (Rietveld, 1984) utilise la régression linéaire pour déterminer les points d'inflexion de la courbe. Dans les années 1980 et 1990, on a essayé d'obtenir de façon automatique une stylisation par mouvements standardisés ('t Hart, 1979 ; Spaai *et al.*, 1993). Le système Momel (Hirst *et al.*, 1993 ; Campione *et al.*, 2000) modélise la courbe de F0 par une fonction de spline quadratique, comme une suite de segments de parabole.

Ces approches reposent donc le plus souvent sur les propriétés mathématiques ou statistiques de la courbe de F0. Les travaux de (Rossi *et al.*, 1981) et (Mertens, 1989) sont exceptionnels en ce sens qu'ils intègrent des seuils de perception.

### 3.2. Stylisation basée sur la perception tonale

(d'Alessandro *et al.*, 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, cf. §3.4.

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.

### 3.3. 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 *et al.*, 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, 1989), mais elles reposent sur les propriétés acoustiques et ne sont pas assez robustes.

La méthode de stylisation décrite a été évaluée (d'Alessandro *et al.*, 1995 ; Mertens *et al.*, 1997) dans des tests perceptifs où les sujets devaient discriminer deux stimuli synthétisés (par la technique PSOLA (Moulines & Charpentier, 1990)), l'un avec les valeurs originales de F<sub>0</sub>, 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<sub>0</sub> : quand le trait épais de la stylisation couvre (et cache) le trait fin du F<sub>0</sub>, la stylisation est identique au F<sub>0</sub> à 1 demi-ton près.

### 3.4. Modifications du modèle dans le prosographe

La stylisation utilisée dans le prosographe s'écarte du modèle original décrit dans (d'Alessandro *et al.*, 1995) sur plusieurs points.

1. Dans le modèle de 1995, une fonction était appliquée à la courbe de F<sub>0</sub> avant la stylisation proprement dite. Sa fonction était de prendre en compte le passé récent (de l'ordre de 50 ms) de la courbe de F<sub>0</sub> 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.

2. Dans son implantation sous forme de script Praat, l'algorithme de stylisation a été entièrement refait. (Le langage de programmation en question ne prévoit pas de variables locales à une procédure, ce qui constitue un obstacle majeur à la mise en œuvre de la récursivité.) Le point d'inflexion majeur dans une courbe entre les points  $(x_1, y_1)$  et  $(x_2, y_2)$  sera défini comme l'instant  $x_t$  pour lequel la différence d'amplitude  $dy$  entre cette courbe et la droite reliant  $(x_1, y_1)$  et  $(x_2, y_2)$  est maximale et dépasse une valeur minimale  $dy_{min}$  égale à 1 ST. Un point d'inflexion ne sera retenu que si la différence de pente pour les parties de part et d'autre du point d'inflexion, soit  $(x_1 .. x_t)$  et  $(x_t .. x_2)$ , dépasse le seuil différentiel de glissando. (Il est cependant

difficile de parler ici d'une pente, puisque la variation en question peut changer au cours de l'intervalle temporel considéré. La pente de la variation est seulement estimée comme l'intervalle mélodique (en ST) entre les valeurs extrêmes, divisée par sa durée.)

Pour le noyau vocalique à styliser, on vérifie d'abord que la variation dépasse le seuil de glissando. Les étapes de segmentation suivantes ne sont appliquées qu'aux intervalles temporels à variation audible. On recherche les points d'inflexion par ordre d'importance ; on repère le plus important  $x_i$  dans l'intervalle  $(x_1 .. x_2)$ , puis le suivant dans l'intervalle  $(x_1 .. x_i)$ , puis celui dans  $(x_i .. x_2)$ , et ainsi de suite, jusqu'à ce qu'il n'y en ait plus. La procédure reprend pour l'intervalle entre d'une part le dernier point d'inflexion retenu, donc celui le plus près de  $x_1$ , par exemple  $x_{i'}$ , et d'autre part le point d'inflexion suivant sur l'axe du temps, soit  $x_i$ , en ainsi de suite. S'il ne reste plus de point d'inflexion dans l'intervalle  $(x_1 .. x_i)$ , la procédure continue pour l'intervalle  $(x_i .. x_2)$ .

Après la segmentation de la courbe aux points d'inflexion majeurs, la stylisation proprement dite est triviale et consiste à remplacer les variations inaudibles par une ligne horizontale à la fréquence du point d'arrivée et les variations audibles par une droite entre leur fréquence de départ et leur fréquence d'arrivée. Pour tout segment autre que le premier, le point de départ coïncide avec le point d'arrivée du segment précédent. L'application du seuil de glissando prend donc la priorité sur la segmentation de la courbe ; c'est ce qui constitue la différence principale par rapport à l'algorithme original de 1995.

#### 4. Application de la stylisation à la transcription automatique de la prosodie

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

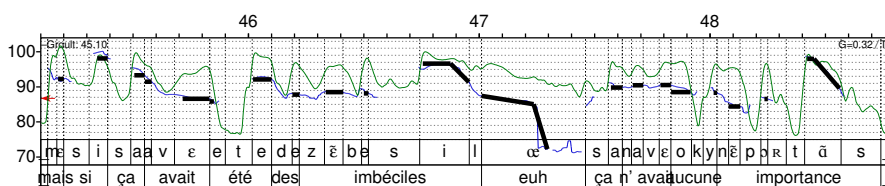
##### 4.1. Délimitation de l'intervalle à styliser

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 requise provient d'un alignement phonétique, donnant, pour chaque son dans le signal, le symbole phonétique et les temps 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, provoquant un saut vers le bas dans le tracé de F0 et dans la stylisation résultante.

Dans une langue comme le français, où le contour mélodique de la syllabe finale de groupe intonatif peut présenter une variation mélodique au cours de la voyelle et des consonnes voisées qui la suivent éventuellement, la variation mélodique perçue peut s'étendre à ces consonnes, surtout si la syllabe est suivie d'une pause (ce qui rejoint les observations de (House, 1995) mentionnées dans la section 2, point 4). C'est le cas des contours d'implication, d'interrogation et de continuation majeure (dans la terminologie de P. Delattre). Pour représenter correctement la mélodie perçue, il serait souhaitable d'inclure dans le domaine du noyau syllabique les consonnes voisées de la coda. À cet effet on adopterait pour les syllabes suivies d'une pause une délimitation modifiée du noyau syllabique. Elle consisterait à inclure dans le noyau syllabique la voyelle et les consonnes de la coda, et d'appliquer comme seuil d'atténuation maximale une valeur de l'ordre de 15 dB. Cette procédure alternative, qui ne demanderait que des modifications restreintes, n'a pas encore été mise en œuvre, parce qu'il n'est pas clair si elle serait justifiée pour des langues autres que le français.

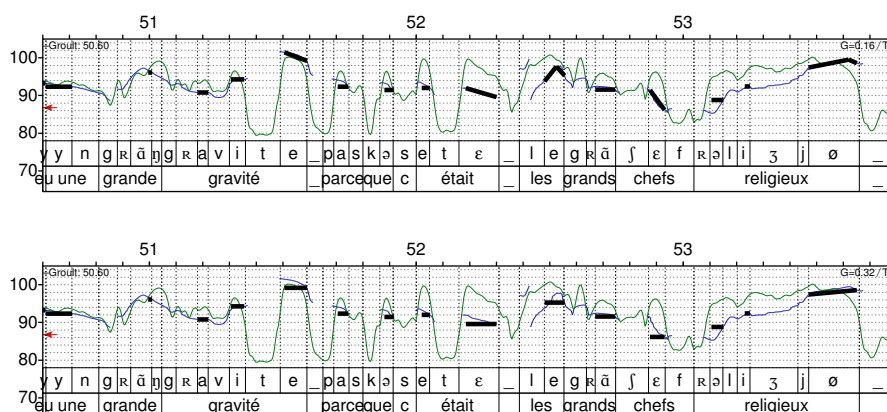
#### 4.2. *Choix des paramètres perceptuels pour la parole continue*

Le seuil de glissando est un des paramètres du modèle de perception tonal et donc de la stylisation. Comme mentionné plus haut, ce seuil dépend de la nature du signal : son pur ou son de parole, son isolé ou parole continue. Alors comment déterminer le seuil de glissando adéquat pour la transcription prosodique ?

Des expériences antérieures (d'Alessandro *et al.*, 1995 ; Mertens *et al.*, 1997) ont indiqué une importante variabilité inter-sujets quant au seuil de glissando. Qui plus est, la capacité d'un auditeur à discriminer des variations mélodiques et à les caractériser évolue dans le temps, comme le montre une longue pratique de l'entraînement auditif chez des étudiants en phonétique. La variabilité importante du jugement des auditeurs complique le choix du seuil à utiliser en transcription automatique. Pour cette raison, nous avons préféré adopter une solution pragmatique qui consiste à essayer de reproduire le jugement de transcrip-teurs expérimentés. De toute façon, ce seuil peut être ajusté dans l'outil de transcription.

La stylisation obtenue pour différentes valeurs du seuil a été comparée de façon systématique avec la transcription manuelle d'un corpus test (le corpus Fayard-Groult, utilisé pour un colloque organisé à 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 du phonéticien averti et a fortiori celles 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 de la décision entre glissando et ton statique. Ces données confirment dès lors 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 entre syllabes successives ou réparties sur plusieurs syllabes.

La figure 6 illustre l'effet du seuil de glissando  $G$  utilisé pour la stylisation. La partie supérieure, qui utilise le seuil bas, retient 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 sans doute 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  $F_0$  à la fois rapide et importante.



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

## 5. Validation

### 5.1. Validation comparative

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 à un colloque sur la prosodie à Genève en septembre 2002, et des corpus R. Barthes (extrait de l'émission "Radioscopie" de J. Chancel, de 17/02/1975, environ 11 min.) et F. Giroud (extrait "Radioscopie" de 15/09/1977, environ 9 min.), pour lesquels (Mertens, 1987) 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  $G = 0.32/T^2$ .

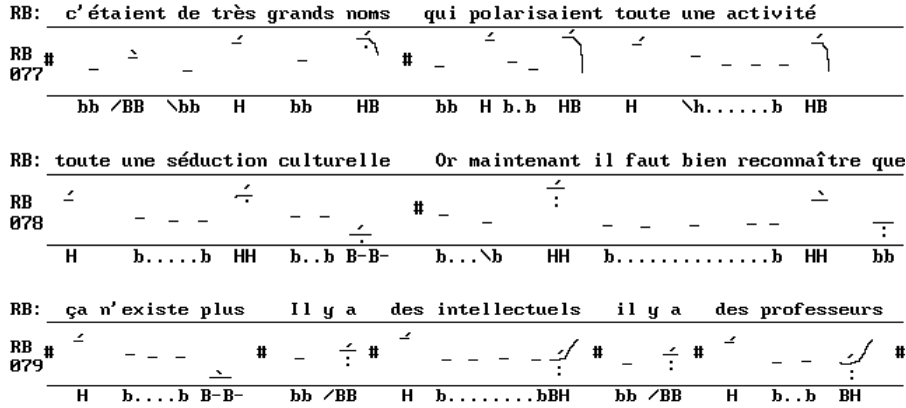


Figure 7. Transcription manuelle tirée du corpus Barthes (Mertens 1987a)

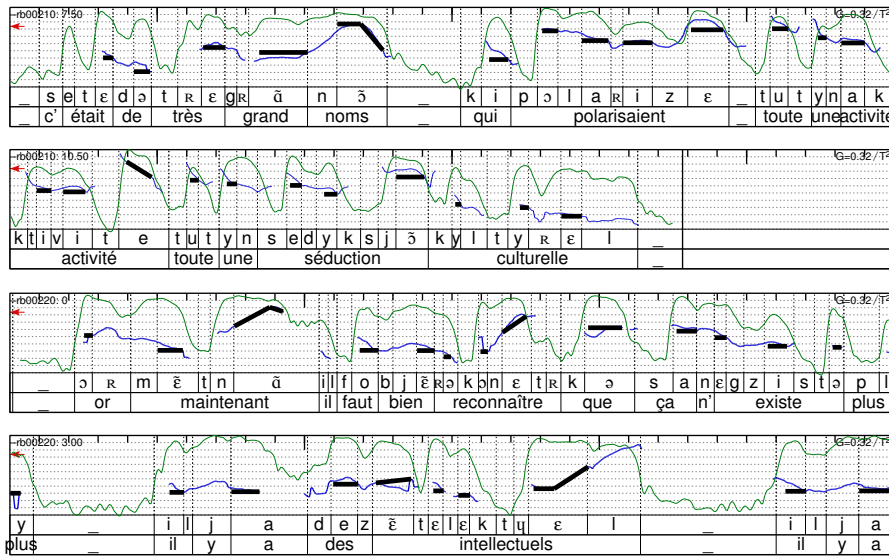


Figure 8. Prosogrammes de l'extrait de la figure 7.

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 en temps énorme.

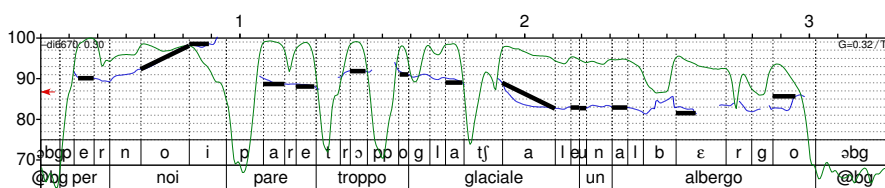
### 5.2. Validation par resynthèse

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 (Moulines & Charpentier, 1990), préserve l'organisation temporelle (durée des segments). Dans la plupart des cas, le signal resynthétisé peut difficilement être distingué du signal original.

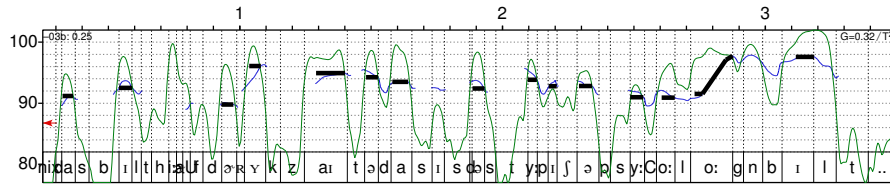
Des fragments resynthétisés sont disponibles sur le site Internet mentionné.

## 6. Utilisation de l'outil pour la transcription de corpus oraux

L'outil de transcription a été mis en œuvre pour obtenir la transcription prosodique de plusieurs corpus de français parlé, d'une durée totale d'environ 30 min, comportant dix locuteurs différents, 4 hommes et 6 femmes. Dans la plupart des cas la segmentation phonétique était disponible ; dans d'autres cas elle a été faite manuellement. Des extraits des transcriptions sont repris sur le site dédié au prosogramme, ainsi que dans (Simon, 2004). Des applications à des corpus d'allemand, de néerlandais et d'italien sont également en cours.

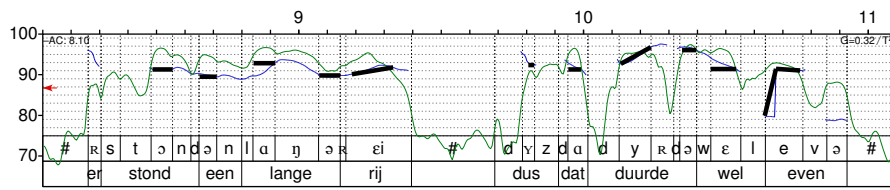


**Figure 9** Prosogramme riche de l'énoncé italien "Per noi pare troppo glaciale, un albergo".



**Figure 10** Prosogramme riche de l'énoncé allemand "das Bild hier auf die Rückseite das ist das typische Psychologenbild".

Plusieurs formats d'affichage ont été prévus en vue des usages envisagés. Le format « large » convient pour les illustrations dans des publications. Le format « compact » prend un minimum de place ; il a été conçu pour l'impression au format A4 de la transcription prosodique d'un corpus entier.



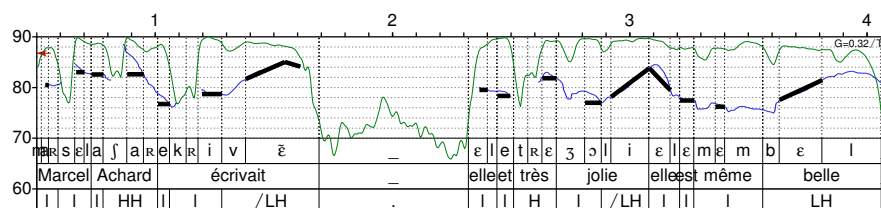
**Figure 11** Prosogramme riche de l'énoncé en néerlandais flamand "Er stond een lange rij, dus dat duurde wel even". Une erreur de détection du fondamental sous forme de saut d'octave apparaît à la syllabe initiale du mot "even".

Le script accepte en entrée des fichiers son isolés ou des ensembles de fichiers. Ceux-ci seront spécifiés par une expression régulière (par exemple « \*.wav ») et par le répertoire où ils sont localisés. Ils seront alors analysés un à un, dans l'ordre alphanumérique des noms de fichier. Cependant, dans les fichiers de sortie, les prosogrammes des différents fichiers se suivent directement pour former une seule transcription continue.

L'utilisateur peut spécifier les temps du début et de la fin de l'extrait souhaité, ainsi que le temps représenté par chaque prosogramme (sa durée, l'intervalle temporel sur l'axe horizontal). Ceci permet de préparer des illustrations qui n'affichent que le fragment souhaité. Pour faciliter l'interprétation des prosogrammes (durée absolue, débit, pente des variations mélodiques...), il est conseillé de garder constante la durée, autour de 3 s par bande.

La segmentation phonétique requise provient d'un fichier au format TextGrid du logiciel Praat. Deux conventions d'encodage sont supportées : Sampa et les « special symbols » de Praat. Une option permet d'afficher les symboles Sampa

comme des symboles de l'alphabet phonétique. Un TextGrid peut contenir *plusieurs couches d'annotation* (appelées « tiers »). L'outil de transcription prosodique peut les afficher dans le prosogramme, sous l'annotation phonétique, et permet à l'utilisateur de sélectionner les couches à afficher.



**Figure 12** Prosogramme comportant plusieurs couches d'annotation.

La réalisation de l'outil de transcription comme un script Praat permet de bénéficier de toutes les fonctionnalités offertes par ce logiciel. Pensons à la correction interactive de l'annotation phonétique, l'écoute interactive du signal, la réorganisation des couches d'annotation à l'intérieur du fichier TextGrid, l'utilisation de la stylisation pour la resynthèse de la transcription prosodique, et ainsi de suite.

De nos jours, il est plus commode dans la *gestion des corpus* de n'utiliser qu'un seul fichier long pour tout l'enregistrement sonore plutôt que d'avoir un nombre élevé de fichiers relativement brefs. Cependant, pour les fichiers longs, il n'est pas toujours possible, ni nécessaire de préparer l'annotation phonétique pour l'ensemble du corpus. Pour cette raison, l'outil de transcription permet de n'analyser qu'un fragment du fichier d'entrée et de limiter le calcul des paramètres au fragment sélectionné. Ceci permet d'associer les avantages de fichiers longs à ceux de fichiers brefs : calcul rapide des résultats et exploitation des résultats.

Dans les usages mentionnés plus haut (illustrations, extraits, corpus entiers), le *fichier de sortie* comprendra le plus souvent l'ensemble des prosogrammes de l'extrait, avec un maximum d'une page A4 (le nombre de prosogrammes par page dépend du mode de visualisation choisi, large ou compact, et du nombre de couches d'annotation affichées). Il est également possible d'obtenir des fichiers de sortie ne comportant qu'un seul prosogramme (une seule bande). Ceci a été prévu en vue de l'intégration de la transcription prosodique dans des bases de données de parole et en vue de leur consultation par le réseau Internet, à l'aide d'un navigateur générique. Dans ces applications, il importe d'assurer l'accès rapide aux prosogrammes et de réduire la quantité de données à transmettre par le réseau.

## 7. Conclusion et perspectives

(Campione *et al.*, 2001: 123) résumant 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, phonétique ou autres, l'annotation phonétique 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 pauses 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.).

Pour certains, l'utilisation du seuil de glissando comme un paramètre dans la stylisation mettrait en cause la « neutralité » du modèle, puisqu'un changement de sa valeur affecte la stylisation. Au contraire, la présence de seuils s'explique par l'objectif même de modéliser la perception auditive. Evidemment, il s'agit d'ajuster ce seuil à sa valeur optimale, obtenue dans des expériences psychoacoustiques pour des stimuli aussi proches que possible de la parole réelle. Refuser l'utilisation de seuils de perception revient en fait à refuser le rôle (ou même l'existence) de la perception. Dans une telle perspective il ne reste plus qu'à utiliser la courbe de fréquence fondamentale, tout en ignorant les nombreux paramètres qui affectent la détection du fondamental : plage des valeurs de fréquence acceptées, scores associés aux candidats, pénalités attribuées aux discontinuités... Vu la multitude de procédés de détection du fondamental, on peut formuler certaines réserves quant à la « neutralité » du paramètre de  $F_0$ .

Deux formats de transcription ont été élaborés. Le format concis ne retient que la stylisation mélodique alignée avec les annotations (phonétique, textuelle, ou autres).

Le format riche prévoit en outre le tracé de  $F_0$  (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 l'ensemble d'un corpus. Dans les deux variantes, on peut afficher jusqu'à cinq couches d'annotation (« tiers »).

L'outil a d'abord été utilisé pour transcrire plusieurs corpus de français parlé, faisant intervenir dix locuteurs (hommes et 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.

Le prosographe est également utilisé en orthophonie (du néerlandais), dans le cadre d'une recherche sur les propriétés mélodiques des voix normales et des voix d'enfants ayant reçu un implant cochléaire.

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. La délimitation du noyau vocalique peut poser problème, par exemple pour les diphtongues en allemand ou en italien, où l'intensité peut varier considérablement au cours de la diphtongue et présente parfois plusieurs pics. A 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. Enfin, des erreurs de détection de la hauteur, en particulier des sauts d'octave, se propagent dans la stylisation et dans la transcription (cf. figure 11) ; ceci vaut pour tout système de transcription basé sur la fréquence fondamentale.

#### Remerciements

Cette recherche a été effectuée en partie dans le cadre du projet plurifacultaire "Prosodie", subventionné par l'Université de Genève, plus particulièrement dans les équipes LATL et « Analyse du discours » dirigées respectivement par E. Wehrli et par E. Roulet. 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, en 1994), A.-C. Simon pour sa contribution à la transcription auditive manuelle du corpus Groult, et P. Boersma et D. Weenink pour la mise à disposition du logiciel Praat.

## 7. Bibliographie

- Alessandro C. d', Rosset S., Rossi J.P., « The pitch of short-duration fundamental frequency glissandos », *J. Acoust. Soc. Am.* 104(4), 1998, p. 2339-2348.
- Alessandro C. d', Mertens P., « Automatic pitch contour stylization using a model of tonal perception », *Computer Speech and Language* 9(3), 1995, p. 257-288.
- Beaugendre F., Hermes D.J., Leenhard G., « Automatic labelling of prosodic events », *IPO-Annual Progress Report* 31, 1996, p. 92-99.
- Campione E., Hirst D., Véronis J., « Stylisation and symbolic coding of F0 : comparison of five models », dans Botinis A., (dir.), *Intonation: Analysis, Modelling and Technology*, Kluwer Academic Publishing, 2000, p. 185-208.
- Campione E., Véronis J., « Etiquetage prosodique semi-automatique des corpus oraux », *Actes TALN 2001*, 2-5 juillet 2001, Tours, p. 123-132.
- Coustenoble H.N., Armstrong L.E., *Studies in French intonation*, Cambridge, Heffer, 1934.
- Geoffrois, E., Extraction robuste de paramètres prosodiques pour la reconnaissance de la parole, Thèse de doctorat, Université Paris XI Orsay, 20 décembre 1995.
- Hart J. 't, « Discriminability of the size of pitch movements in speech », *I.P.O. Annual Progress Report* 9, 1974, p. 56-63.
- Hart J. 't, « Psychoacoustic backgrounds of pitch contour stylisation », *I.P.O. Annual Progress Report* 11, 1976, p. 11-19.
- Hart J. 't, « Explorations in automatic stylization of F0 curves », *I.P.O. Annual Progress Report* 14, 1979, p. 61-65.
- Hart J. 't, « Differential sensitivity to pitch distance, particularly in speech », *J. Acoust. Soc. Am.* 69(3), 1981, p. 811-821.
- Hart J. 't, Collier R., Cohen, A., *A perceptual study of intonation*, Cambridge: Cambridge Univ. Press, 1990.
- Hermes D.J., « Vowel-onset detection », *I.P.O. Annual Progress Report* 22, 1987, p. 15-24.
- Hermes D.J., van Gestel J.C., « The frequency scale of speech intonation », *J. Acoust. Soc. Am.* 90(1), 1991, p. 97-102.
- Hirst D., Espesser, R., « Automatic Modelling of Fundamental Frequency Using a Quadratic Spline Function », *Travaux de l'Institut de Phonétique d'Aix-en-Provence* 15, 1993, p. 75-85.
- House D., *Tonal Perception in Speech*. Lund: Lund Univ. Press, 1990.
- House D., « The influence of silence on perceiving the preceding tonal contour », *Proc. Int. Congr. Phonetic Sciences* 13, Stockholm, vol. 1, 1995, p. 122-125.
- Mertens P., L'intonation du français. De la description linguistique à la reconnaissance automatique. Thèse de doctorat, Université de Leuven, 1987.

- Mertens P., « Automatic segmentation of speech into syllables », *Proceedings of the European Conference on Speech Technology*, Edinburgh, vol. II, 1987, p. 9-12.
- Mertens P., « Automatic recognition of intonation in French and Dutch », *Proceedings Eurospeech 89*, Paris, 1989, vol. 1, p. 46-50.
- Mertens P., d'Alessandro Ch., « Pitch contour stylization using a tonal perception model », *Proc. Int. Congr. Phonetic Sciences 13*, vol. 4, 1995, p. 228-231.
- Mertens P., Beaugendre F., d'Alessandro Ch., « Comparing approaches to pitch contour stylization for speech synthesis », dans Santen J.P.H. van, Sproat R.W., Olive J.P., Hirschberg J. (dir.), *Progress in Speech Synthesis*. N.Y.: Springer Verlag. 1997, p. 347-363.
- Mertens P., Goldman J.-P., Wehrli E., Gaudinat A., « La synthèse de l'intonation à partir de structures syntaxiques riches », *Traitement Automatique des Langues* 42(1), 2001, p. 145-192.
- Moulines E., Charpentier F., « Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones », *Speech Communication* 9, 1990, p. 453-467.
- Rietveld A.C.M., Syllaben, klemtonen en de automatische detectie van beklemtoonde syllaben in het Nederlands, Thèse de doctorat, Université de Nijmegen, 1984.
- Rossi M., « Le seuil de glissando ou seuil de perception des variations tonales pour la parole », *Phonetica* 23, 1971, p. 1-33.
- Rossi M., « La perception des glissandos descendants dans les contours prosodiques », *Phonetica* 35 (1), 1978, p. 11-40 .
- Rossi M., « The perception of non-repetitive intensity glides on vowels », *Journal of Phonetics* 6 (1), 1978, p. 9-18.
- Rossi M., « Interactions of intensity glides and frequency glissandos », *Language & Speech* 21, 1978, p. 384-396.
- Rossi M., Di Cristo A., Hirst D., Martin Ph., Nishinuma Y., *L'intonation. De l'acoustique à la sémantique*, Paris: Klincksieck, 1981, 364 pp.
- Spaai G.W.G., Storm A., Derksen A.S., Hermes D.J., Gigi E.F., « An Intonation Meter for teaching intonation to profoundly deaf persons », *IPO Manuscript* no. 968, 1993.
- Simon A.C., *La structuration prosodique du discours en français. Une approche multidimensionnelle et expérimentale*, Bern, Peter Lang, 2003.
- Taylor P., « Automatic recognition of intonation from F0 contours using the rise / fall / connection model », *Proceedings Eurospeech 1993*, 1993, p. 789-792.