

Evaluation de modèles d'étiquetage automatique de l'intonation

Estelle Campione, Emmanuel Flachaire, Daniel Hirst, Jean Véronis

Laboratoire Parole et Langage,
Université de Provence & CNRS
29, Av. Robert Schuman, 13621 Aix-en-Provence Cedex 1, France
Tel. : +33 4 42 95 36 33, Fax : +33 4 42 59 50 96
E-mail: Estelle.Campione@lpl.univ-aix.fr

ABSTRACT

In this paper we discuss various problems related to the automatic labelling of F_0 contours through the evaluation of five models of increasing complexity. The models were tested on a French corpus and the best two models were subsequently evaluated for Italian (twenty speakers for a total of 90 minutes of speech). Although there is still some room for improvement, the best two models ensure a quite satisfactory resynthesis of the original F_0 contours (mean quadratic error of about one semi-tone on either absolute F_0 values or relative pitch intervals).

1. INTRODUCTION

L'étiquetage prosodique de grands corpus serait extrêmement utile pour l'étude de la parole en général, ainsi que pour des tâches d'ingénierie telles que la génération d'une prosodie de qualité en synthèse [Ver97] ou bien de l'utilisation de la prosodie en reconnaissance. Les quelques corpus qui existent (par exemple [Ost95]) ont été étiquetés manuellement par des experts. Outre le caractère lent et difficile de cette tâche, son caractère éminemment subjectif diminue la fiabilité des résultats, ou impose le recours à des contre-expertises qui augmentent les coûts. Il serait donc extrêmement intéressant de disposer de systèmes d'étiquetage automatique permettant de s'affranchir de la phase d'intervention manuelle, ou de la réduire à une simple phase de vérification et de correction. Toutefois, on ne dispose pas, à l'heure actuelle, d'étiqueteur prosodique automatique fiable, même pour les systèmes de transcription les plus utilisés, tels que ToBI, bien que des recherches soient en cours ([Bla96], [Dus97]).

Nous discutons dans cette communication de divers problèmes liés à l'étiquetage automatique de l'intonation, à travers la présentation de cinq modèles de complexité croissante. Les cinq modèles ont été testés sur un corpus français, et les deux meilleurs modèles ont été ensuite évalués sur l'italien (20 locuteurs et 90 minutes de parole au total). Bien que perfectibles sur divers aspects, les deux meilleurs modèles proposés permettent une bonne régénération des contours intonatifs originaux (erreur quadratique

moyenne inférieure au demi-ton sur les hauteurs absolues ou les intervalles mélodiques).

2. PROBLÉMATIQUE

On considère souvent que la F_0 est la combinaison d'une composante macroprosodique qui reflète le choix d'un patron intonatif par le locuteur, et d'une composante microprosodique qui ne résulte pas d'un choix du locuteur (abaissement des constrictives, etc.) [DiC86]. De nombreuses études se sont attachées, depuis les années 60, à factoriser ces deux composantes, et à extraire le plus automatiquement possible l'information macroprosodique pertinente du signal de parole. Cette extraction peut être décomposée en deux étapes :

- *stylisation*, c'est-à-dire remplacement de la courbe de F_0 par une fonction numérique plus simple conservant la même information macroprosodique ;
- *codage symbolique*, c'est-à-dire représentation par une suite de symboles constituant une discrétisation de la courbe stylisée.

La première étape est parfois qualifiée de *close-copy stylisation* ([DeP79]), dans la mesure où il est souhaitable que des énoncés re-synthétisés, dans lequel la F_0 serait remplacée par la F_0 stylisée, ne soient pas perceptuellement distincts de l'original. Cette notion de *close copy* peut être étendue au codage symbolique : il serait souhaitable que la discrétisation effectuée ne perde pas d'information pertinente en ce qui concerne la macroprosodie, et que la F_0 régénérée à partir du codage symbolique (à l'aide d'un modèle adéquat) ne se distingue pas non plus perceptuellement de la courbe stylisée et de l'original. On aurait ainsi un système de stylisation et de codage totalement réversible, qui serait du plus grand intérêt pour le codage automatique de grands corpus de parole.

La stylisation a fait l'objet de nombreux travaux (par exemple : [tHa90], [Hir93], [Tay94], [DA195], etc.), et on peut dire que la technique en est maîtrisée de façon satisfaisante. Divers systèmes de codage symbolique ont également été proposés, dont le plus connu est sans doute ToBI [Sil92]. Toutefois, l'automatisation et la réversibilité (*close copy*) sont beaucoup moins

avancées pour le codage symbolique que pour la stylisation.

3. ASPECTS COMMUNS AUX 5 MODÈLES

Les cinq modèles testés utilisent tous la même méthode de stylisation automatique, MOMEL (MODélisation MELodique), a été proposée par Hirst et Espesser [Hir93] (voir aussi [Hir98a]). Contrairement à la plupart des autres méthodes de stylisation, MOMEL utilise des fonctions splines quadratiques (c'est-à-dire des arcs de parabole), lisses en tout point, et non des segments des droites, qui introduisent des angles vifs consécutifs. De plus, les segments non voisés sont interpolés et la courbe stylisée ne présente donc aucune discontinuité. Les courbes splines quadratiques utilisées peuvent être représentées par une suite de *points-cibles* correspondant aux seuls changements significatifs (passages par zéro de la tangente), à partir desquels le codage symbolique peut être calculé.

Le plus connu des systèmes de codage symbolique est ToBI [Sil92], utilisé avec succès par de nombreux auteurs sur l'anglais. Toutefois, ToBI présente l'inconvénient d'être un système spécifique à une langue : son adaptation à d'autres langues, bien que possible, nécessite un effort assez important. Par ailleurs, l'étiquetage ToBI repose sur des jugements linguistiques pris par un expert, et est difficilement automatisable (bien que des tentatives existent dans ce sens : [Bla96]). Enfin, la régénération d'une F_0 proche de l'original à partir de ToBI est délicate.

Les modèles présentés dans cette étude sont inspirés par le système de codage symbolique INTSINT (INternational Transcription System for INTonation) proposé par Hirst et Di Cristo [Hir98b]. Au contraire de ToBI, qui encode des événements de nature linguistique, INTSINT est un codage purement formel de la courbe macroprosodique (cette idée est aussi utilisée par Taylor [Tay93], dans le système HLCCB). Chaque symbole permet de coder un point cible, soit de façon absolue, soit par rapport au registre du locuteur ou bien de façon relative par rapport aux points environnants. Ce type de codage formel peut être considéré comme un premier niveau de codage, pouvant servir de base, le cas échéant, à des systèmes de plus haut niveau tels que ToBI.

Comme dans le système INTSINT, tous les modèles testés comportent deux symboles absolus T (*top*) et B (*bottom*) qui servent à coder les points extrêmes, ainsi qu'un symbole relatif S (*same*) indiquant une absence de mouvement mélodique :

- les points cibles dépassant un seuil t_T sont codés T , ceux au-dessous d'un seuil t_B sont codés B (t_T et t_B sont choisis de façon que 5% des points

soient codés T et autant soient codés B en supposant une distribution normale) ;

- les points cibles situés à moins de 2.5% (en demi-tons ou ST) du point précédent sont codés S .

Les modèles diffèrent par le codage relatif des mouvements ascendants et descendants (voir sous-section suivante). Le premier point, s'il n'est pas codé T ou B est codé par rapport à la moyenne des fréquences des points-cibles du segment de parole modélisé.

La régénération s'effectue à partir d'un point fictif situé avant le début de l'énoncé et de fréquence égale à la moyenne. Les points suivants sont générés de la façon suivante :

- les points T correspondent à une fréquence F_T et les points B à une fréquence F_B respectivement déterminées par la moyenne des points au-dessus de t_T et au-dessous de t_B dans une distribution normale ;
- les points S gardent la fréquence du point précédent ;
- les fréquences des autres points relatifs sont calculés par régression linéaire à partir du point précédent, à l'aide de coefficients de régression estimés pour chaque symbole sur le corpus à modéliser (tous locuteurs confondus en valeurs centrées réduites pour une langue donnée).

4. SPÉCIFICITÉS

4.1. Modèle HL

Le premier modèle testé est un modèle minimal qui comporte seulement deux symboles en plus des deux symboles T , B et S communs à tous les modèles :

- H : montant ;
- L : descendant.

4.2. Modèle Config

Le second modèle testé est un modèle dont les symboles utilisés représentent la configuration locale des points cibles:

- H : pic ;
- U : plateau montant (*upstepped*) ;
- D : plateau descendant (*downstepped*) ;
- L : vallée.

4.3. Modèle Ampli2

Le modèle *Ampli2* est basé sur des symboles représentant les amplitudes des mouvements ascendants et descendants :

- H_2 : ascendant d'amplitude supérieure à un seuil a ;
- H_1 : ascendant d'amplitude inférieure à a ;
- L_1 : descendant d'amplitude inférieure à a ;
- L_2 : descendant d'amplitude supérieure à a .

Par essais systématiques, nous avons déterminé que les meilleurs résultats de ce modèle sont obtenus lorsque a correspond à 40% de $t_T - t_B$.

4.4. Modèle Ampli3

Le modèle *Ampli3* est basée sur la même idée mais utilise trois niveaux d'amplitude déterminés par deux seuils a_1 et a_2 :

- H_3 : ascendant d'amplitude supérieure à a_2 ;
- H_2 : ascendant d'amplitude entre a_1 et a_2 ;
- H_1 : ascendant d'amplitude inférieure à a_1 ;
- L_1 : descendant d'amplitude inférieure à a_1 ;
- L_2 : descendant d'amplitude entre a_1 et a_2 ;
- L_3 : descendant d'amplitude supérieure à a_2 .

Les meilleurs résultats de ce modèle sont obtenus lorsque a_1 et a_2 correspondent respectivement à 25% et 50% de $t_T - t_B$.

4.5. Modèle Niveaux

Le modèle *Niveaux* utilise un principe différent. La bande centrale entre les deux seuils t_T et t_B est divisée en trois bandes, correspondant chacune à 30 % des points dans une distribution normale : *G* (Grave), *M* (Médium), *A* (Aigu). Le codage des points relatifs se fait en associant la direction par rapport au point précédent et la bande dans laquelle se trouve le point :

- H_A : ascendant, aboutissant dans l'aigu ;
- H_M : ascendant, aboutissant dans le médium ;
- H_G : ascendant, aboutissant dans le grave ;
- L_A : descendant, aboutissant dans l'aigu ;
- L_M : descendant, aboutissant dans le médium ;
- L_G : descendant, aboutissant dans le grave.

5. ÉVALUATION

5.1. Corpus

Les cinq modèles ont d'abord été évalués sur le français, puis les deux meilleurs modèles ont ensuite été testés sur l'italien. Le corpus utilisé est une partie de la base de données EUROM1 [Cha95], comprenant 35 minutes de parole pour le français et 54 pour l'italien. Les phrases de ce corpus sont regroupées en passages de 5 phrases prononcés par 10 locuteurs par langue (5 hommes et 5 femmes).

La totalité du corpus a été stylisé automatiquement à l'aide de l'algorithme MOMEL, puis vérifié et corrigé manuellement par des experts. Le corpus contient

6329 points cibles pour le français et 9804 pour l'italien. Les fréquences de points cibles ont été converties en demi tons (ST). L'écart type varie de 2.79 ST à 4.18 ST pour les locuteurs français et de 2.78 ST à 4.44 ST pour les locuteurs italiens. Les points initiaux des énoncés d'un locuteur ont une moyenne très proche de la moyenne de l'ensemble des points cibles de ce locuteur (écart moyen de -0.88 ST en français et 0.34 ST en italien). La distribution peut être approximée de façon grossière par une loi normale.

5.2. Méthode

Les courbes régénérées ont été comparées avec les courbes stylisées vérifiées par des experts. Comme il a été souvent noté dans la littérature, la mesure de la similarité entre deux courbes intonatives n'est pas simple, et en tout état de cause devrait faire intervenir des tests perceptifs, rarement possibles à grande échelle. Nous utiliserons comme mesures :

- e_{abs} : erreur quadratique moyenne sur les fréquences des points-cibles (en ST) ;
- e_{rel} : erreur quadratique moyenne sur l'amplitude des mouvements mélodiques (en ST) ;
- p_{abs} : pourcentage de points-cibles à moins de 2 ST de la fréquence originale ;
- p_{rel} : pourcentage de mouvements mélodiques à moins de 2 ST de l'amplitude originale.

5.3. Résultats

Les résultats sur le français sont récapitulés par les figures 1 et 2. La valeur de e_{abs} est minimale pour le modèle *Niveaux* (0,96 ST), celle de e_{rel} est minimale pour le modèle *Ampli3* (1,09 ST). Le pourcentage p_{abs} est maximal pour le modèle *Niveaux* (96,0%), et le pourcentage p_{rel} est maximal pour le modèle *Ampli3* (93,1%).

Les résultats obtenus sur l'italien en ce qui concerne les deux meilleurs modèles, seuls testés sur cette langue, sont analogues.

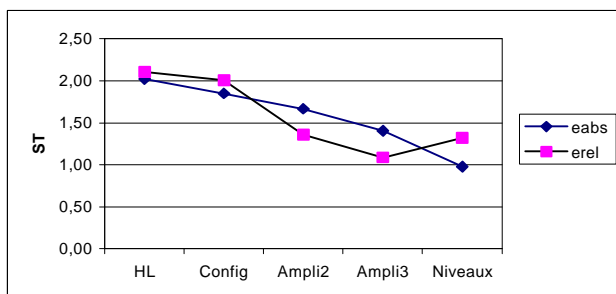
6. DISCUSSION

Comme il est prévisible, les résultats s'améliorent avec le nombre de symboles utilisés. On note cependant les points suivants :

- Le modèle *HL* est clairement trop fruste pour être utilisable de façon pratique.
- A nombre de symboles égaux, le modèle *Ampli2* surpasse le modèle *Config*. Il ne semble donc pas que la configuration (pic, plateau) soit porteuse de suffisamment d'information pour une bonne régénération. Au demeurant, la configuration peut être dérivée de façon triviale du modèle *Ampli2* par l'examen de la séquence de symboles.

- Enfin, les deux meilleurs modèles, *Ampli3* et *Niveaux*, de nombre de symboles égaux, ont un comportement opposé : *Ampli3* permet une bonne restitution des mouvements mélodiques, mais une moins bonne restitution des hauteurs absolues, alors que *Niveaux* se comporte de façon inverse. On note également que les résultats sont similaires en français et en italien, ce qui laisse penser que les modèles concernés sont relativement indépendants de la langue.

Une discrimination entre les deux meilleurs modèles demanderait des tests perceptifs de façon à déterminer s'il est préférable d'optimiser la restitution des hauteurs absolues ou celle des intervalles. Ces modèles sont de plus susceptibles d'améliorations diverses. Il semblerait en particulier que des symboles additionnels soient nécessaires pour coder les points les plus extrêmes du registre du locuteur, qui sont les



points les plus mal régénérés dans tous les modèles.

Figure 1: Erreur quadratique moyenne (français)

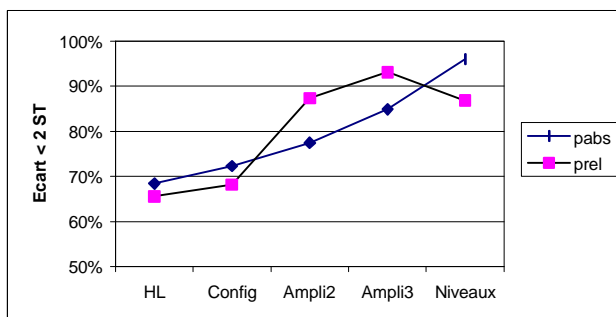


Figure 2: Pourcentage d'écarts <2ST (français)

REMERCIEMENTS

Les auteurs remercient Robert Espesser pour son aide technique, Corine Astésano et Fabienne Courtois pour leur travail de correction du corpus.

BIBLIOGRAPHIE

- [Bla96] Black, A., Hunt, A. (1996). Generating F_0 contours from ToBI labels using a linear regression, *Proc. ICSLP'96*, Philadelphia.
- [Cha95] Chan, D., et al. (1995). EUROM- A Spoken Language Resource for the EU. *Proceedings of Eurospeech'95*. Madrid, 1, 867-870.

- [DAI95] D'Alessandro, C., Mertens, P. (1995). Automatic pitch contour stylization using a model of tonal perception, *Computer Speech and Language*, 9, 257-288.
- [DeP79] De Pijper, JR. (1979). Close-copy stylisation of British English intonation contour, *IPO Annual Progress Report* 14, 66-71.
- [DiC86] Di Cristo, A. Hirst, D.J. (1986). Modelling French micromelody : analysis and synthesis. *Phonetica*, 43, 1/3, 11-30.
- [Dus97] Dusterhoff, K., Black, A. (1997). Generating F_0 Contours for Speech Synthesis Using the Tilt Intonation Theory, *Proceedings of an ESCA Tutorial and Research Workshop on Intonation : Theory, Models and Applications*, September 1997, Athens, 107-110.
- [Hir93] Hirst, D., Espesser, R. (1993). Automatic Modelling of Fundamental Frequency using a quadratic spline function, *Travaux de l'Institut de Phonétique d'Aix-en-Provence*, 15, 75-85.
- [Hir98a] Hirst, D., Di Cristo, A., Espesser, R. (forthcoming). Levels of representation and levels of analysis for the description of intonation systems. In Horne, M. (Ed.), *Prosody: Theory and Experiment*, Dordrecht: Kluwer Academic Publishers.
- [Hir98b] Hirst, D., Di Cristo, A. (in press) A survey of intonation systems. in Hirst, D., Di Cristo, A. (Eds) *Intonation Systems: a Survey of Twenty Languages*. Cambridge : Cambridge University Press, 1-44.
- [Ost95] Ostendor M.F., Price P. J., Shattuck-Hufnagel S. (1995). The Boston University Radio News Corpus, , *Technical Report No. ECS-95-001*, Boston University.
- [Ros72] Rossi, M., Chafcouloff, M. (1972). Les niveaux intonatifs. *Travaux de l'Institut de Phonétique d'Aix*, 1,167-176.
- [Sil92] Silverman, K., et al. (1992). ToBI: a standard for labelling English prosody. *Proc. ICSLP'92*, 2, 867-870, Banff, Canada.
- [Tay93] Taylor, P. (1993). Automatic Recognition of Intonation from F_0 -contours using the Rise/Fall/Connection Model, *Proceedings of Eurospeech'93*, Berlin, 2, 789-792.
- [Tay94] Taylor, P. (1994). The Rise/Fall/ Connection Model of Intonation, *Speech Communication*, 15:1&2, 169-186.
- [tHa90] t'Hart, J., Collier, R., Cohen, A. (1990). A perceptual study of intonation : an

experimental-phonetic approach to speech melody, Cambridge Univ. Press.

- [Ver97] Véronis *et al.* (1997). A stochastic model of intonation for French text-to-speech synthesis, *5th European Conference on Speech Communication and Technology, Eurospeech'97*, Rhodes (Greece), September 1997.