Stylisation automatique de la fréquence fondamentale: une évaluation multilingue

Corine Astésano, Robert Espesser, Daniel Hirst

Laboratoire Parole et Langage, URA 261 CNRS, Institut de Phonétique d'Aix,
Université de Provence, 29, av. Robert Schuman, 13621 Aix-en Provence Cedex 1

Joaquim Llisterri

Departament de Filologia
Edifici B, Universitat Autonoma de Barcelona
08193 Bellaterra, Barcelona

courrier-e: *Corine.Astesano@lpl.univ-aix.fr*

Summary: This paper presents a multilingual evaluation of the prosodic tools developped in Aix (Laboratoire Parole et Langage) in the framework of the MULTEXT project. These tools allow an automatic stylisation of the fundamental frequency curve as a sequence of target points (MOMEL), as well as a symbolic coding with the INTSINT system. The evaluation has been carried out on 5 languages (English, French, German, Spanish and Swedish) from the EUROM1 database. For each language, the corpus consists of 40 passages composed of thematically linked sentences in order to provide a linguistic and prosodic coherence adapted to the task. After stylisation of the f0, a perceptual comparison between the resynthesis and the original sentence is carried out, and the target points are adjusted, if necessary, in order to obtain a melodic contour perceptually acceptable to a native speaker.

1. INTRODUCTION

Le projet européen MULTEXT (Véronis & al., 1994) a pour but de contribuer au développement de logiciels largement diffusables pour la manipulation et l'analyse du texte et de la parole multilingues. Une partie du projet essaie d'intégrer certains outils et méthodes du traitement du langage naturel et du traitement de la parole par l'étude de phénomènes à l'intersection des deux domaines, en particulier la prosodie, qui entretient des relations complexes avec la morphologie et la syntaxe.

Le laboratoire Parole et Langage a développé des outils d'étiquetage automatique de la prosodie. Ces outils permettent la stylisation automatique de la F0 sous forme de points cibles (MOMEL, Hirst et Espesser, 1993), ainsi qu'un codage symbolique de ces points cibles dans le système INTSINT (Hirst et Di Cristo, à paraître).

Nous décrivons ici les résultats de l'évaluation de ces outils sur cinq langues (anglais, allemand, espagnol, français et suédois). Dans chacune des langues, la F0 de 40 passages de 5 phrases issus du corpus EUROM1 a été automatiquement stylisée, et la resynthèse de ces passages a été ensuite évaluée par des experts pour chacune des langues.

2. OUTILS

L'algorithme de modélisation automatique (MOMEL) permet la représentation de la fréquence fondamentale par une séquence de points cibles constituée par des couples de valeurs <F0, temps>. Les points cibles correspondent aux variations locales pertinentes de la courbe mélodique et permettent, interpolés par une fonction de type spline quadratique, de retrouver le profil suprasegmental caractérisant globalement l'intonation. Un outil d'analyse/resynthèse (technique PSOLA, Hamon & al. 1989), utilisant les points cibles ainsi détectés, permet la resynthèse de la courbe originale à partir de la courbe modélisée. Ce procédé est utilisé pour la validation perceptive de la stylisation automatique de la f0.

3. DESCRIPTION DU CORPUS

Une des tâches du projet MULTEXT (Llisterri, 1996) consistait à fournir un étiquetage prosodique de la base de données de parole multilingue EUROM1 (projet ESPRIT SAM 2589). Une description complète du corpus est disponible dans Sherwood & Fuller (1992).

Dans le cadre du projet MULTEXT, 5 langues ont été retenues (anglais, allemand, espagnol, français, et suédois). Les passages composés de phrases thématiquement liées entre elles, issus du 'Few Speaker Set' d' EUROM1, constituent le matériel de base de notre étude, car ils proposent une cohérence linguistique et prosodique. Pour chaque langue, 10 locuteurs se répartissent la lecture de 40 passages (5 phrases par passage), ce

qui correspond à une centaine de passages par langue (500 phrases). Le matériel d'étude a été réduit ici de façon à ce que la taille du corpus soit proportionnelle à l'effort à fournir dans le cadre du projet Multext (200 phrases par langue). Le matériel d'étude totalise donc 1000 phrases et 20 locuteurs, également répartis entre hommes et femmes.

4. METHODOLOGIE

La même méthodologie a été appliquée pour les 5 langues, et comportait 4 phases principales:

~ détection de la f0 (combinaison de 3 méthodes de calcul: fonction peigne, amdf, autocorrelation).
~ détection des points cibles (modélisation automatique de la f0, MOMEL)
~ validation perceptive de la modélisation: comparaison auditive entre la courbe originale de f0 et la resynthèse obtenue par la modélisation. L'éventuelle correction de points cibles intervenait jusqu'à ce que la resynthèse soit jugée acceptable pour un natif.
~ alignement manuel des mots orthographiques et des syllabes accentuées.

5. RESULTATS

Les résultats présentés concernent la phase de validation perceptive de la stylisation automatique de la f0 pour les 5 langues. Pour cette tâche, deux analyses différentes ont été menées pour l'espagnol, le français et l'allemand d'une part (analyse quantitative des problèmes de la stylisation automatique), et pour le suédois et l'anglais d'autre part (analyse qualitative de la stylisation automatique).

L'écart moyen (toutes langues confondues) entre la courbe stylisée et la courbe originale de f0 (moyenne de l'écart absolu terme à terme entre les valeurs stylisées et les valeurs originales) est de 5,44 % pour une durée totale de 3 h 45 mn de parole.

5.1. Résultats de l'évaluation quantitative des erreurs de Momel

Pour l'espagnol, l'allemand et le français, trois catégories de problèmes relatifs à la stylisation automatique ont été définies:

~ les **points cibles** (PC) **manquants** qui ont dû être rajoutés manuellement de façon à ce que la resynthèse à partir de la courbe stylisée soit acceptable pour un natif. Ces PC ont été relevés dans trois contextes: en position initiale, en position médiane, et en position finale de phrase où l'on a distingué les contours finaux montants et descendants.
~ les **points cibles redondants** pouvant être supprimés sans incidence sur la synthèse.
~ les **points cibles déplacés** soit sur un axe horizontal, soit lorsqu'ils étaient détectés trop haut ou trop bas par rapport à la courbe originale.

Langues	Durée (mn) des segments analysés (40 passages)	PC manquants	PC déplacés	PC redondants	Total
Français	14	4,9	0,4	0,075	5,47
Allemand	14	5,1	4,5	4,78	14,28
Espagnol	13	4,2	0,55	0,16	4,81

Tableau des pourcentages de points cibles erronés par catégories

La disparité observée entre l'allemand et les 2 autres langues est due à une stratégie de correction manuelle des PC propre à l'expert, qui s'est attaché à éliminer un maximum de PC non nécessaires à la resynthèse ("bruit"). L'élimination de PC entraîne le plus souvent le déplacement de PC adjacents. Dans l'ensemble, la catégorie des PC manquants est la plus importante et représente 60% des cibles erronées toutes catégories confondues. Ces PC manquants peuvent induire des erreurs de modélisation importantes, en particulier sur les contours mélodiques finaux ascendants: 30% de ces contours sont mal modélisés. A un degré moindre, le même constat peut être fait pour les contours mélodiques initiaux. Il est intéressant de noter que ces erreurs de modélisation interviennent majoritairement au voisinage des pauses. La prise en compte de ce phénomène diminuerait considérablement les erreurs de détection des PC, et améliorerait d'autant la stylisation. Les experts constatent néanmoins que les quelques erreurs de détection n'handicapent en rien l'intelligibilité des phrases synthétisées.

5.2. Résultats de l'évaluation qualitative de Momel

Pour le suédois et l'anglais, les évaluateurs se sont attachés à décrire les écarts de la modélisation automatique par rapport à la courbe originale en terme de changements de modalité ou changements de sens induits par la synthèse. Pour le suédois notamment (Aasa et Strangert, 1996), certaines erreurs de détection des PC entraînaient un déplacement de l'accent lexical implicant un changement de sens ou donnant une impression d'accent dialectal. Pour l'anglais néanmoins, ces erreurs n'ont donné lieu qu'à des changements de modalité.

6. CONCLUSION ET DIRECTIONS FUTURES

MOMEL offre une bonne représentation des événements pertinents de la courbe de f0 quelle que soit la langue. Il ressort de cette évaluation multilingue qu'un traitement approprié des pauses augmenterait sensiblement les performances de MOMEL.

L'extension de cette méthode à l'ensemble du corpus EUROM1, ainsi que le codage symbolique des points cibles dans le système INTSINT constituent la deuxième phase du projet MULTEXT. A terme, l'objectif est la création de grandes bases de données

prosodiques dans le but, par exemple, de permettre l'apprentissage de modèles stochastiques, tels que celui décrit par Courtois & al. dans ces actes.

Références:

Aasa, A., Strangert, E. (1996), 'Prosodic analysis of Swedish within the Multext project', *Fonetik 96,* Nässlingen 29-31 mai 1996.

Courtois, F., Di Cristo, Ph., Lagrue, B., Véronis, J., (1997), 'Un modèle stochastique des contours prosodiques en français pour la synthèse à partir de textes', *dans ces actes.*

Hamon, C., Moulines, E., Charpentier, F., (1989), ' A diphone system based on time-domain prosodic modifications of speech', *Proc. ICASSP 89, pp. 238-241.*

Hirst, D., Di Cristo, A. (à paraître), 'A survey of intonation systems'. In Hirst, D., Di Cristo, A. (eds), *Intonation Systems: a survey of twenty languages.*Cambridge University Press.

Hirst, D., Espesser, R. (1993), 'Automatic modelling of fundamental frequency'. *Travaux de l'Institut de Phonétique d'Aix*, 15, pp.71-85.

Llisterri, J. (1996), *Prosody Tools Efficiency and Failures,* WP 4 Corpus, T4.6 Speech Markup and Validation, Final, 15 October 1996, MULTEXT-LRE Project 62-050.

Sherwood, T., Fuller, H. (1992), *Guide to EUROM1 Speech Database.* Doc. No. SAM-NPL-102, Final, 21 April 1992. ESPRIT PROJECT 2589.

Véronis, J., Hirst, D., Espesser, R., Ide, N. (1994), 'NL and speech in the Multext project'. *AAAI'94 Workshop on Integration of Natural Language and Speech*, pp.72-78.