CONSTITUTION SEMI-AUTOMATIQUE DE LEXIQUES DE CONTOURS PROSODIQUES POUR LA SYNTHESE A PARTIR DU TEXTE

V. Aubergé et M. Contini

ICP -Université Stendhal, UA 368-
38000 Grenoble - France

ABSTRACT

Each application of text-to-speech synthesis requires a specific prosodic strategy (depending on the context). Furthermore, the elaboration of these strategies is tightly linked to the analysis capabilities of the automatic text processing, which is then to reproduce the prosody of the application-specific corpus. A complete system was elaborated to handle at the same time the textual data and the prosodic data of the corpus. The successive processing modules of such a system are : automatic extraction, visualization and averaging of prosodic contours.
The next step of this work is the constitution of different prosodic modules, sutiable for varying linguistic situations.

1. INTRODUCTION

Lorsque l'on aborde les technologies vocales, et en particulier la synthèse à partir du texte, on s'aperçoit que la qualité segmentale de la voix synthétique est devenue suffisante pour que la qualité de l'intonation synthétique soit maintenant un problème prioritaire, puisque l'intonation est directement reliable à l'intelligibilité et pas simplement à des critères de naturalité ou d'agrément.
Dans un système artificiel, l'entrée est un texte, c'est-à-dire une projection sémiotique du langage. Pour transformer ce message symbolique en message parlé, on peut retrouver dans l'écrit certaines des structures du langage mais on ne sait rien d'éventuels modèles psycholinguistiques du locuteur. C'est pourquoi, afin d'amélorer la qualité prosodique d'un systèrne de synthèse on propose ici une méthodologie, basée sur l'analyse inductive de corpus [1] & [2] qui a permis la conception d'un module de génération de l'intonation pour la synthèse.
Le corpus est construit *a priori* selon des critères linguistiques que l'on qualifiera d'hypothético-déductifs. Dans un premier temps, des hypothèses sont émises sur des points d'ancrage des structures intonatives autour d'indices déductibles de l'écrit. A partir de ces indices linguistiques est construit un corpus symbolique, qui devient un corpus oral après enregistrement. Ensuite, toutes les données du corpus décomposées et classifiées sont moyennées et constituent une base de contours paramétrés par des indices linguistiques. Si on est capable de bien choisir les liens de coïncidence entre les structures de l'écrit (hypothèses a priori) et celles de l'intonation, nous pouvons espérer que les représentants des classes intonatives permettront de reconstruire une structure anthropophonique. La génération de l'intonation se résume ensuite à une application hiérarchisée de ces contours pendant la phase de synthèse.

2. LE CORPUS

Le choix du corpus est fondamental. Il doit être le résultat d'une démarche logiquement corrélée avec la méthode d'analyse qui doit ensuite s'appliquer au corpus. Sa qualité de représentativité est déterminante quelle que soit la finesse de l'analyse qu'on en fera. Puisque l'entrée du synthétiseur est un texte, le choix des indices pertinents est en pratique fortement limité par les capacités des analyses textuelles automatiques avec lesquelles il est nécessaire d'entretenir des relations étroites.
L'hypothèse la plus forte retenue a priori est celle de "rendez-vous" structurels entre intonation et syntaxe [3]. Cette relation sera d'autant plus émergente du corpus que le locuteur est mis dans une situation de lecture de phrases isolées non marquées sémantiquement. Parallèlement, il est possible, avec les analyseurs du CRISS de produire une analyse morpho-syntaxique automatique de la phrase [4].
Le corpus a été construit pour être analysé hiérarchiquement par niveaux syntaxiques successifs : l'hypothèse étant que non seulement l'intonation coïncide avec la syntaxe par morceaux, mais surtout qu'il est possible d'associer une unité intonative caractéristique à chaque niveau considéré.
Le niveau maximal représenté dans notre corpus est la phrase. Les niveaux inférieurs étudiés sont dans l'ordre la proposition, le syntagme, l'unité lexicale et enfin la syllabe.
A chacun de ces niveaux sont étudiées l'influence de la longueur (en nombre de syllabes) de l'unité considérée et sa position syntagmatique à l'intérieur du ou des niveaux supérieurs. Chaque unité est caractérisée par des attributs spécifiques qu'on suppose distinctifs pour les unités intonatives : e.g. la modalité pour la phrase, la fonction de dépendance pour la proposition, la nature ou la fonction syntaxique pour le syntagme, son organisation en constituants... Le niveau sémantique est abordé par l'utilisation de traits lexicaux.
A chaque patron linguistique produit, on choisit un représentant (i.e. un énoncé de phrase textuel) qui respecte des contraintes phonétiques afin de garantir un étiquetage correct du corpus acoustique incident.
Un premier corpus de 164 phrases a été enregistré selon une ergonomie qui, par expérience, laisse espérer une bonne cohérence interne dans la stratégie prosodique du locuteur. Il est à noter que ce locuteur est celui du dictionnaire de polysons qui constitue la méthode de synthèse du système sur lequel a été testé cette méthodologie.

3. L'ANALYSE DU CORPUS

Les paramètres physiques choisis sont classiquement Fo et durée. Le paramètre durée est bien sûr délicat à acquérir puisque s'y projettent aussi bien les structures segmentales que suprasegmentales. Quant à Fo, il a été montré (et nous avons pu le vérifier a posteriori) que l'acquisition de trois points par voyelle (début, extremum, fin) est suffisante pour restituer l'évolution suprasegmentale de ce paramètre pour le français.
Il a été obtenu ainsi une base de données simplifiées, où sont associés le corpus symbolique étiqueté (par les paramètres linguistiques selon lesquels il a été construit), et le codage de Fo et durée. Un ensemble de logiciels de gestion du corpus manipulent et traitent les données en fonction des étiquettes linguistiques qu'elles contiennent. Des outils de visualisation des paramètres acoustiques permettent d'observer les données intonatives manipulées, avec diverses possibilités de normalisation,de myennage et de superposition des contours. La méthodologie analytique du traitement passe à la fois par la classification des contours homogènes et par une analyse contrastive de ces contours :

- Pour décider de la qualité d'une structure du texte à coïncider avec une structure intonative, toutes les structures correspondantes dans le corpus sont rassemblées en une même classe. Un contour-moyen est conservé dans la base en association avec les paramètres linguistiques qui spécifient la structure textuelle de cette classe.

- L'autre méthode d'investigation systématiquement utilisée avant la classification puis ensuite comme vérification sur les contours-moyens calculés est une méthode contrastive par paire minimale. Lorsqu'un paramètre linguistique est supposé pertinent, alors dans le corpus sont choisis deux représentants pour lesquels ce paramètre varie, toute chose étant égale par ailleurs. Si la paire minimale est effective alors l'ensemble des paramètres (les paramètres étudié et contextuels) seront spécifiques d'une classe de contours.

Voici en exemple quelques extraits du corpus symbolique à partir desquels seront comparés les contours intonatifs

correspondant aux éléments notés entre "/" :

paramètre étudié : position respective Nom-Adjectif (construction du groupe)
paramètres contextuels : Groupe Nominal (GN), début de phrase assertive, 4 syllabes
/ Ce beau passant / chantait.
/ Ce passant fou / chantait.

/ Ce fantastique passant / chantait.
/ Ce passant fantastique / chantait.

paramètre étudié : longueur
paramètres contextuels : adjectif, dans la structure GN construit selon le modèle de la phrase 2, début de phrase assertive.

/ Ce beau passant / chantait.
(1 syllabe)
/ Ce petit passant / chantait.
(2 syllabes)
/ Ce fantastique passant / chantait.
(3 syllabes)

paramètre étudié : fonction de dépendance de proposition
paramètres contextuels : 7 syllabes, fin de phrase assertive.

Je vois ces enfants ; / ils jouent avec un balai /.
Quand je vois ces enfants, / ils jouent avec un balai /.
Je vois ces enfants / quand ils jouent avec un balai /.

Avec une telle méthode, il est à chaque instant possible d'affiner l'analyse en intégrant de nouveaux paramètres. La cohérence avec les capacités des analyses textuelles automatiques est ainsi assurée par une simple mise à jour de la base des contours.

4. LA BASE DE CONTOURS

Au niveau de la phrase, l'unité intonative conservée est la ligne de déclinaison pour chaque longueur de phrase et pour chaque modalité. La cohérence à l'intérieur de chaque classe est vérifiée (écart-type < 1/4 ton). Les lignes-moyennes sont, elles aussi, très cohérentes (écart-type <1/4 ton pour les assertives et 1/3 pour les interrogatives), c'est donc la pente qui varie avec la longueur de la phrase, les valeurs des bornes restent constantes. Des variations locales entre chaque classe sont utilisées en synthèse pour une rupture de la monotonie.
Au niveau de la proposition, la fonction de dépendance a un intérêt distinctif évident. C'est également une ligne (attaque, finale) qui est calculée pour chaque classe paramétrée par la fonction de dépendance, la position dans la phrase et la longueur.
Au niveau inférieur, celui du syntagme, intervient pour chaque groupe un paramètre booléen (fin,non fin) de phrase assertive. Les contours d'un même groupe en position fin et non fin sont similaires jusqu'à l'avant dernière syllabe. Quant à la dernière syllabe, elle présente une pente négative de valeur à peu près identique quelque soit le groupe. Plutôt que de calculer une règle de transformation d'un contour de non fin à fin, ont été stockés les deux types de contours. L'unité la plus finement représentée est le groupe nominal (GN). On retrouve dans ce corpus la limite de quatre syllabes pour le groupe intonatif. Les constituants ne sont pas particularisés à l'intérieur d'un GN de moins de quatre ou cinq syllabes (quelque soit la construction du GN, son contour est similaire) et deviennent des unités intonatives dans les GN plus grands (les contours du GN s'opposent selon sa construction). Pourtant, si l'on observe le comportement de l'adjectif ou du nom, lorsque sa longueur varie à l'intérieur d'un même GN, son évolution reste très cohérente et tout à fait indépendante de la longueur du GN. Les contours-moyens des GN sont stockés, paramétrés par une variable de position dans la phrase (fin, non fin), par la situation par rapport au verbe et par une indication booléenne sur la fonction syntaxique du GN en valeur de complément circonstant ou pas. Pour les GN de un, deux, trois et quatre syllabes, ce contour sera maximal ; pour les GN de plus de quatre syllabes, les contours des constituants interviendront dans la reconstruction de la structure pendant la phase de génération.
Les groupes verbaux ont été classifiés selon leur appartenance ou non aux verbes outils, leur longueur et leur position dans la phrase.
Une étude restreinte des adverbes autonomes a donné des contours particularisés par leur position dans la phrase et par rapport au verbe.

5. CONCLUSION

Cette démarche a pour principe de formuler des hypothèses linguistiques les plus faibles possibles (mais alors taille du corpus ?) et de laisser les données "s'auto-classifier". Le rôle de l'expert est isolé et une étape ultérieure consistera à le substituer par un système formel.
Des tests de validation de l'intonation générée sont en cours. Les premiers résultats sont encourageants : l'intonation obtenue est contrastée et de plus elle paraît effectivement se rapprocher de l'intonation de notre locuteur. Il s'est avéré que sa stratégie s'est bien organisée selon les hypothèses posées à la construction du corpus, sans doute en partie parce que ce locuteur a été placé en situation non spontanée de lecture. En dehors de tout contexte discursif, il n'a eu comme support formel que la syntaxe de phrase.
Cette première expérience a pu valider cette méthodologie pour la synthèse. On peut facilement imaginer que la même méthode de constitution et d'analyse de corpus soit appliquée à d'autres langues avec l'ensemble des outils mis en place.
L'intonation est un facteur de qualité essentiel de la synthèse vocale si elle devient spécifique par exemple d'une situation de messagerie, de dialogue homme-machine, de bureautique multi-media ou de commandes à distance, chacune de ces applications étant caractérisable par une base de contours. Cette méthode sera sera toujours limitée par le choix de paramètres linguistiques pertinents, en fonction du contenu intonatif du corpus et de la capacité des analyses linguistiues automatiques, mais on peut supposer que plus les spécifications linguistiques seront précises, moins les structures intonatives seront étendues.

6. REFERENCES

[1] CONTINI M. & PROFILI O. (1987), *"Génération automatique de schémas macroprosodiques en italien, à partir d'un texte écrit."*, 16 èmes JEP - GALF, 245-248.
[2] EMERARD F. & BENOIT C. (1987), *"De la production à l'extraction, l'état d'un chantier."*, 16 èmes JEP - GALF, 224-228.
[3] ROSSI M. DI CRISTO A. HIRST D. MARTIN P. & NISHINUMA Y. (1981), *"L'intonation, de l'acoustique à la sémantique"*, Klincksieck Ed., Paris.
[4] ROUAULT J. (1988), *"Linguistique automatique, applications documantaires"*, Sciences pour la Communication, Peter Lang, Berne.