TALN 2010, Montréal, 19–23 juillet 2010

Expressive : Génération automatique de parole expressive à partir de données non linguistiques

Olivier Blanc[1] Noémi Boubel[1] Jean-Philippe Goldman[2] Sophie Roekhaut[3]
Anne Catherine Simon[2] Cédrick Fairon[1] Richard Beaufort[1]
(1) CENTAL, Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgique
(2) Valibel, Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgique
(3) TCTS Lab, Université de Mons, 7000 Mons, Belgique

Résumé. Nous présentons Expressive, un système de génération de parole expressive à partir de données non linguistiques. Ce système est composé de deux outils distincts : *Taittingen*, un générateur automatique de textes d'une grande variété lexico-syntaxique produits à partir d'une représentation conceptuelle du discours, et *StyloPhone*, un système de synthèse vocale multi-styles qui s'attache à rendre le discours produit attractif et naturel en proposant différents styles vocaux.

Abstract. We present Expressive, a system that converts non-linguistic data into expressive speech. This system is made of two distinct parts : *Taittingen*, a natural language generation tool able to produce lexically and syntactically rich texts from a discourse abstract representation, and *StyloPhone*, a text-to-speech synthesis system that proposes varying speaking styles, to make the speech sound both more attractive and natural.

Mots-clés : Génération de texte, synthèse vocale, expressivité.

Keywords: Natural language generation, text-to-speech synthesis, expressiveness.

1 Introduction

La démocratisation de l'accès à une masse d'informations toujours croissante, combinée à l'émergence de nouvelles formes de médias, amène à inventer de nouvelles formes de transmission de l'information, capables notamment de tenir compte du profil de l'utilisateur. Le projet Expressive [1] s'inscrit dans ces considérations. Son objectif est de développer un outil permettant de transmettre de l'information sous la forme d'une parole expressive adaptée aux besoins de l'utilisateur. À cette fin, Expressive combine deux outils de traitement automatique des langues : un système de génération automatique de texte et un système de synthèse de la parole à partir du texte.

Le module de génération produit un texte français compréhensible et bien structuré à partir d'une représentation conceptuelle et non linguistique d'un discours. Le système gère les variations aux niveaux de la structure du discours, des formes syntaxiques et du lexique et est ainsi capable de produire de nombreux

1. Projet de recherche subsidié par la Région wallonne de Belgique, convention no. 616422.
Voir : `http://cental.fltr.ucl.ac.be/team/projects/expressive/`

textes différents à partir des mêmes données de départ, évitant ainsi la monotonie souvent problématique dans les textes automatiquement générés. Le texte obtenu est ensuite traité par le synthétiseur vocal, qui convertit le texte en parole en veillant à respecter un style expressif prédéfini, afin de rendre le discours plus vivant et naturel tout en conservant son intelligibilité.

La chaîne de traitement complète, en cours de finalisation, est actuellement testée sur une base de données biographiques dont dispose le CENTAL (Kevers & Fairon, 2007). L'idée, dans le cadre de ce test, est de générer automatiquement la biographie vocale d'une personnalité à partir des informations disponibles dans la base. Les deux modules sont adaptables plus largement, ensemble ou séparément, à de nombreux autres domaines et applications. Ces composants ne sont pas actuellement mis à la disposition de la communauté ; les modalités de distribution seront décidées d'ici la fin du projet, prévue pour décembre 2010.

2 Taittingen : système de génération automatique de textes

Développé en Python, *Taittingen* est un système de génération complet : il produit un texte français à partir d'une représentation conceptuelle et non linguistique d'un discours, proche de celle utilisée dans la *Théorie de la structure rhétorique* (Mann & Thompson, 1988). Son architecture modulaire reprend les grandes lignes des architectures en *pipeline* classique (Reiter & Dale, 2000) et a été décrite en détails dans (Blanc & Boubel, 2009).

Le système est actuellement testé en s'appuyant sur les données d'une base de données biographiques. Cette base est régulièrement alimentée par un système d'extraction d'informations, qui analyse les articles de presse en ligne et répertorie automatiquement un certain nombre d'évènements biographiques typiques : naissance, mort, professions exercées, diplômes obtenus, etc. Notre application permet ainsi de générer un court texte sur la vie d'une personne répertoriée dans la base.

Taittingen est capable de produire un très grand nombre de textes différents à partir des mêmes données de base. Actuellement, le système est prévu de sorte que le texte généré pour une entrée donnée soit différent à chaque relance du programme. À terme, l'objectif est de choisir parmi ces variantes en fonction des données disponibles et des desiderata de l'utilisateur : demande d'un texte bref, d'informations particulières à mettre en avant, etc. L'étude en cours sur un corpus de biographies nous a permis de dégager des sources de variabilité à différents niveaux.

Durant la génération, le système utilise un dictionnaire de transfert décrivant les possibilités de traduction de chaque concept propre au domaine de la biographie vers une représentation lexico-syntaxique. La variété est ainsi rendue possible principalement par enrichissement manuel de ces ressources linguistiques. À titre d'exemple, les évènements *éducations* et *professions* regroupent chacun une trentaine de patrons lexico-syntaxiques différents permettant d'exprimer la même idée. De la même manière, le système exploite une ressource de connecteurs discursifs, ce qui permet de faire varier la façon d'exprimer les relations qui lient les segments d'un discours.

Enfin, nous travaillons actuellement à l'ajout de variantes au niveau de la planification du discours. L'idée est d'adapter la structure générale du texte généré en fonction des informations disponibles dans la base et du profil de l'auditeur ou du lecteur. D'après nos observations, les textes biographiques ne présentent pas une énorme liberté stylistique et suivent généralement une structure assez figée. Ici, quatre plans du discours ont été envisagés : textes courts, textes structurés par ordre chronologique, textes structurés par thématiques et textes sous forme de listes énumératives.

3 StyloPhone : Système de synthèse multi-styles

Le système de synthèse à la base du projet est eLite (Beaufort & Ruelle, 2006) qui, dans sa dernière version, produit de la parole par sélection d'unités non-uniformes (NUU, *Non-Uniform Units*). Dans ce type de système, la prosodie originale a été conservée. Si cette approche facilite l'obtention d'une courbe mélodique relativement naturelle, elle complique, par contre, l'intégration de variations mélodiques, parce que le style de la synthèse est fortement dépendant de la voix enregistrée.

Développé autour d'eLite, *StyloPhone* est un synthétiseur multi-styles : il cherche à introduire des variations de style tout en conservant le naturel de la voix de synthèse. Trois styles vocaux sont actuellement modélisés : journalistique, politique et conversationnel.
Pour obtenir un style vocal donné, *StyloPhone* modifie les valeurs de certains paramètres prosodiques de la parole. Les paramètres prosodiques exploités appartiennent à différentes catégories :
– les variations phonologiques : pourcentage de e muets finaux prononcés et pourcentage de pauses pleines (respirations, hésitations) ;
– les durées : durée des syllabes, durée des pauses, nombre de syllabes entre deux pauses ;
– l'intonation et l'accentuation : variation de la F0, nombre de syllabes proéminentes, nombre de syllabes avec accent initial, nombre de syllabes avec accent final.
Selon les paramètres, les modifications sont réalisées soit au cœur d'eLite, soit en post-traitement. Les modifications réalisées au cœur d'eLite sont les variations phonologiques et la modélisation des durées. Elles sont effectuées *avant* la sélection des unités, afin de conserver une courbe prosodique naturelle. Les modifications appliquées en post-traitement sur le signal de parole produit par eLite concernent l'intonation et l'accentuation (Roekhaut *et al.*, 2010). Elles sont réalisées à l'aide d'un script Praat (Boersma & Weenink, 2010) lancé en ligne de commande. Au cours de ce traitement, en fonction des variations observées dans un style donné, certaines syllabes initialement non-proéminentes peuvent devenir proéminentes.

Afin de déterminer les modifications à réaliser, les valeurs des paramètres prosodiques ont été étudiées sur un corpus de parole expressive, et comparées aux valeurs produites par eLite en synthèse neutre. Le corpus de parole que nous avons utilisé est C-PROM, qui contient l'enregistrement de plusieurs locuteurs et de plusieurs styles (Avanzi *et al.*, 2010). L'extraction des valeurs des paramètres est réalisée à l'aide des outils ProsoReport (Goldman *et al.*, 2007a) et ProsoProm (Goldman *et al.*, 2007b).

Les modifications réalisées permettent de générer 3 styles de synthèse qui reproduisent les paramètres observés dans chaque style original. Cependant, ces modifications ne sont pas encore totalement satisfaisantes. Afin de se rapprocher le plus possible d'une production orale naturelle et réaliste, il nous semble par exemple utile d'orienter les recherches vers la modélisation du style d'un seul locuteur, parce que les variations entre plusieurs locuteurs au sein d'un même style sont trop importantes.
Nous constatons également une perte de naturel due aux modifications en post-traitement. Pour pallier cet inconvénient, nous avons l'intention de favoriser les modifications au cœur du système, avant la sélection des unités de parole.

4 Démonstration

Dans le cadre de la démonstration, nous présenterons une application qui intègre, sous une même interface graphique, le module de génération automatique de textes et le module de synthèse vocale. Le

programme permet d'accéder aux informations présentes dans la base de données biographiques. Pour chaque personne référencée dans cette base, l'interface présente différentes variantes possibles de courtes biographies sur sa vie et permet d'écouter les discours correspondants oralisés, dans les trois styles prédéfinis. Parmi ceux-ci, le style qui semble actuellement le plus approprié pour présenter les biographies est le style journalistique. Nous envisageons d'étudier un autre style de parole qui nous semble plus pertinent pour la présentation de ce type de texte : l'exposé oral.

La présentation du système se fera sur les machines apportées par les participants et ne nécessitera pas de moyens techniques particuliers.

Références

AVANZI M., SIMON A.-C., GOLDMAN J. & AUCHLIN A. (2010). C-PROM : Un corpus de français parlé annoté pour l'étude des proéminences. In *Proceedings of JEP*, Mons, Belgium. http ://sites.google.com/site/corpusprom.

BEAUFORT R. & RUELLE A. (2006). eLite : système de synthèse de la parole à orientation linguistique. In *Proceedings of JEP*, p. 509–512.

BLANC O. & BOUBEL N. (2009). Taittingen : A Biographical Text Generation System. In Z. VETULANI, Ed., *Proceedings of 4th Language & Technology Conference*, p. 375–379.

BOERSMA P. & WEENINK D. (2010). *Praat : doing phonetics by computer (Version 5.1.24)*. http ://www.praat.org.

GOLDMAN, J.-PH. AUCHLIN A., SIMON A.-C. & AVANZI M. (2007a). Phonostylographe : un outil de description prosodique. Comparaison du style radiophonique et lu. *Nouveaux cahiers de linguistique française*, **28**, 219–237.

GOLDMAN J.-P., AVANZI M., SIMON A.-C., LACHERET A. & AUCHLIN A. (2007b). A methodology for the automatic detection of perceived prominent syllables in spoken French. In *Proceedings of Interspeech 2007*, p. 98–101.

KEVERS L. & FAIRON C. (2007). Vers une base de connaissances biographiques : extraction d'information et ontologies. In *Actes des 7èmes Journées Francophones Extraction et Gestion des Connaissances*, volume / of *Revue des Nouvelles Technologies de l'Information - Série Extraction et gestion des connaissances*, p. 373–378.

MANN W. C. & THOMPSON S. A. (1988). Rhetorical structure theory : Toward a functional theory of text organization. *Text*, **8**(3), 243–281.

REITER E. & DALE R. (2000). *Building Natural Language Generation Systems*. Cambridge.

ROEKHAUT S., GOLDMAN J.-P. & SIMON A.-C. (2010). A Model for Varying Speaking Style in TTS systems. In *Proceedings of Speech Prosody 2010*, Chicago, Illinois, USA.