eLite : système de synthèse de la parole à orientation linguistique

Richard Beaufort, Alain Ruelle

Multitel ASBL
Avenue Copernic 1, 7000 Mons, Belgique
{beaufort,ruelle}@multitel.be
http://www.multitel.be/TTS

ABSTRACT

eLite is the Text-to-Speech synthesis system developed by the TTS-NLP group of Multitel ASBL. The creation of eLite has been an opportunity for the group to carry out and integrate further research on all domains of text-to-speech synthesis, like morphological analysis, syntactic desambiguation and non-uniform units selection. This paper presents the general features and techniques of the system.

1. INTRODUCTION

En synthèse de la parole à partir du texte, la génération du signal de parole n'est pas directement réalisée à partir du texte, mais à partir d'une représentation phonétique et prosodique de celui-ci. Cependant, parce que la langue écrite regorge d'ambiguïtés linguistiques, la génération de la représentation phonétique et prosodique doit elle-même être précédée d'une phase de désambiguïsation du texte. C'est conscient de l'importance d'une analyse linguistique fiable que le groupe TTS-NLP de Multitel ASBL a développé eLite, prononcé [i l a j t], dont le nom signifie ***E**nhanced, **LI**nguistically-based **TE**xt-to-speech synthesis system*.
En ce qui concerne l'étape de génération du signal de parole, les premières versions d'eLite ont intégré le synthétiseur MBROLA, basé sur le principe de concaténation d'unités de parole pré-enregistrées et prosodiquement neutres. La dernière version d'eLite intègre le dernier état de l'art en synthèse, la sélection et la concaténation d'unités non-uniformes : la sélection est réalisée par l'algorithme LiONS, la concaténation, par TP-MBROLA.

Commencé en septembre 2001 et conçu initialement afin de fournir à l'équipe une plateforme complète de test pour de nouveaux algorithmes utiles à la synthèse, eLite est rapidement devenu un logiciel stable, robuste et rapide, dont des versions de démonstration sont disponibles sur le site du groupe (http://www.multitel.be/TTS).

Cet article présente l'architecture générale d'eLite ainsi que les différents modules du processus de synthèse.

2. ARCHITECTURE DU SYSTÈME

Unité linguistique. La totalité de l'architecture d'eLite repose sur une notion fondamentale, celle d'*unité linguistique*. Une unité linguistique, dans eLite, est *un mot ou une séquence de mots et de symboles formant un tout*. L'unité linguistique de base est évidemment le mot, dans le sens de *séquence de caractères alphabétiques comprise entre deux espaces*, l'espace pouvant être un ou plusieurs blancs, retours à la ligne ou signes de ponctuation. Une unité linguistique peut également correspondre à un mot composé dont les constituants sont séparés par un ou plusieurs blancs ou par un tiret. Enfin, l'unité linguistique peut être une *unité de sens*, comme les adresses URL, les numéros de téléphone ou les nombres et unités de mesure.

Modules et Structure de données. eLite (cf. fig. 1) reçoit en entrée un texte et produit en sortie la parole correspondante. Le système se divise en 3 modules principaux :

1. Le NLP (*Natural Language Processing*), qui gère le traitement du langage naturel.
2. La Sélection, qui choisit les unités de parole.
3. Le DSP (*Digital Signal Processing*), qui concatène les unités de parole choisies et produit le signal voulu.

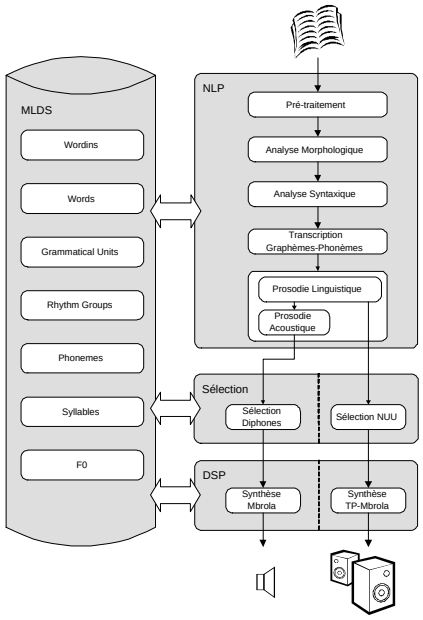

FIG. 1: Architecture d'eLite

Le module NLP se divise lui-même en 5 étapes. Le pré-traitement, l'analyse morphologique et l'analyse syntaxique désambiguïsent le texte. La conversion graphèmes-phonèmes associe une séquence de phonèmes à chaque mot, et gère les phénomènes phonétiques aux frontières des mots. La génération de la prosodie génère des informations prosodiques de type linguistique et

acoustique. Notons que la prosodie acoustique n'est nécessaire que dans le cas d'une synthèse via MBROLA.

Les 3 modules communiquent au travers d'une structure de données multicouches, la MLDS (*Multi-Layers Data Structure*), inspirée de celle proposée par le projet Festival de l'université d'Edinbourg [3]. Notre MLDS se compose de 7 couches :

1. *Wordin* : ce sont les unités linguistiques détectables par le pré-traitement, telles que les URLs, téléphones, etc. Notons qu'un mot composé ne sera par détecté comme un seul Wordin, mais comme plusieurs.
2. *Word* : ce sont les mots au sens défini précédemment. Un Word peut donc être une partie de Wordin, ou correspondre à un Wordin complet. Chaque Word possède une liste de natures possibles.
3. *Grammatical Unit* : une Grammatical Unit est l'analyse syntaxique d'une unité linguistique. Une Grammatical Unit correspond donc généralement à un Wordin. Cependant, dans le cas des mots composés, une Grammatical Unit correspondra à plusieurs Wordins.
4. *Rhythm Group* : ce sont les groupes de souffle d'une phrase, entre lesquels une pause peut survenir. Un Rhythm Group englobe généralement plusieurs Grammatical Units.
5. *Phoneme* : les phonèmes des mots du texte. Un phonème est toujours lié à un Word.
6. *Syllable* : les syllabes, constituées de phonèmes.
7. *F0* : les fréquences fondamentales attribuées aux phonèmes voisés.

Niveaux d'analyse. L'originalité d'eLite est qu'il présente deux niveaux d'analyse syntaxique, grâce à la présence simultanée d'informations syntaxiques dans la couche *Grammatical Unit* et dans les natures de la couche *Word* (Cf. tab. 1).
La couche *Grammatical Unit* donne une vue macroscopique de la phrase, puisqu'elle ne propose une analyse syntaxique que pour les unités linguistiques. Ceci sera particulièrement utile lors de l'analyse syntaxique : de la sorte, la syntaxe n'a pas à s'embarrasser de détails comme, par exemple, de la présence de symboles dans une URL.
Les listes de natures de la couche *Word* fournissent une analyse détaillée de chaque mot. Ceci sera entre autres nécessaire lors de la phase de conversion graphèmes-phonèmes. Il sera par exemple très utile de savoir que l'URL \textit{www.president.fr} contient le nom *président* et non le verbe, afin de générer la phonétisation correcte.

TAB. 1: Words, Natures, Grammatical Units

Word	Natures	Gram. Unit
pommes	NOUN	NOUN
de	PREP	
terre	NOUN	

Langues. Actuellement, eLite traite le français et l'anglais. Néanmoins, le système est ouvert à d'autres langues. En effet, toutes les données nécessaires aux différents processus ont été externalisées. L'ajout d'une langue dans eLite revient donc à concevoir les bases de données qui y correspondent. eLite est cependant actuellement limité aux langues dites *romanes* (français, espagnol, italien, etc.) et *germaniques* (anglais, allemand, néerlandais, etc.), parce que les étapes de l'analyse restent intrinsèquement liées à la structure de ces langues. Des langues telles que le turc, l'arabe ou le chinois ne sont donc pas encore gérables dans eLite.

3. NLP

Le but du NLP est de fournir aux modules suivants une représentation phonétique et prosodique du texte à synthétiser. Toutefois, les ambiguïtés présentes dans la langue écrite nécessitent de commencer par désambiguïser le texte, ce qui est réalisé en 3 étapes : pré-traitement, analyse morphologique et analyse syntaxique.

Pré-traitement. Le rôle du pré-processeur est de diviser le texte en Wordins et de supprimer les caractères parasites (espaces, caractères vides de sens). Les Wordins générés correspondent à des unités linguistiques, sauf dans le cas des mots composés (Cf. *Analyse morphologique*).
Les unités linguistiques à détecter sont modélisées à l'aide d'expressions régulières compilées sous la forme d'une machine à états finis, chargée par le pré-processeur. Les unités linguistiques reconnues sont :

- *Mot* : mangera, ils, TCTS
- *Ponctuation* : . ! ? ; : - . . .
- *URL* : www.cuisiner.fr/index.php?x=10, 10.108.55.9, john.smith@foo.co.uk
- *Date* : 01/02/2002, 30/06/75
- *Heure* : 8h 10min 30s, 10:45
- *Téléphone* : 010/24.38.97, +33 3 27 33 34 54
- *Nombres* : 10.000.000, -10.045,43e-43
- *Montants* : $40000, -10.000 EUR, +43,433.76 USD
- *Mesures* : 25 km/h, 10.343,45 N/m^2
- *Acronymes* : A.S.B.L., T.C.T.S., O.T.A.N.

Lorsqu'un Wordin est détecté, le pré-processeur le segmente en Word à l'aide d'une autre machine à états finis, également générée à partir d'expressions régulières. Par exemple, une URL sera segmentée en symboles (:, ., //), acronymes (http, ftp, www, etc.) et mots (Cf. tab. 2).
Le pré-processeur effectue également une détection sommaire de la mise en page du document. Celle-ci se borne pour l'instant à repérer les titres (chiffres numériques ou romains suivis d'un point en début de ligne), les énumérations (points, astérisques ou tirets en début de ligne) et les paragaphes (multiples sauts de lignes), ce qui permet d'identifier des fins de phrase non marquées par un symbole de ponctuation et d'insérer des pauses de longueur adaptée dans la prononciation du texte.

TAB. 2: Segmentation Wordin → Words

Wordins	Words	
http://www.fortis.be	http	ACRONYM
	:	SYMBOL
	//	SYMBOL
	www	ACRONYM
	.	SYMBOL
	fortis	NOUN
	.	SYMBOL
	be	ACRONYM

Analyse morphologique. Cette analyse s'effectue en deux étapes. Dans un premier temps, elle attribue à chaque Word une liste de natures possibles. Dans un second temps, elle constitue la couche Grammatical Unit.

Pour attribuer une liste de natures à un Word donné, l'analyseur applique les règles suivantes dans l'ordre, et s'interrompt dès qu'une règle produit au moins une analyse :

1. *Wordins spéciaux* :les Wordins qui ne sont pas de type « Mot » peuvent présenter des caractères spéciaux qui seront analysés différemment en fonction du type de Wordin. Par exemple, un point (.) sera considéré comme un symbole s'il appartient à une URL ou à un Téléphone, mais sera analysé comme une ponctuation forte si le Wordin est de type Ponctuation.
2. *Test flexionnel* : cette analyse est tout-à-fait classique. L'analyseur cherche dans le mot des flexions potentielles, et les remplace par les flexions normalisées correspondantes. Si la forme normalisée appartient au dictionnaire de lemme, l'analyse est conservée.
3. *Test de réaccentuation* : l'analyseur cherche les caractères qui pourraient avoir été désaccentués. Par exemple, le mot *eleve* peut être la forme désaccentuée de *élevé* ou *élève*. Chaque forme réaccentuée subit le traitement décrit au point 2. Ceci est particulièrement pertinent pour les URLs, toujours désaccentuées.
4. *Catégories ouvertes* : en l'absence de correcteur orthographique, l'analyseur doit conclure que le mot inconnu est probablement un néologisme. Pour cette raison, il lui attribue la liste des catégories ouvertes aux néologismes. En français, il s'agit de *nom*, *verbe*, *adjectif* et *adverbe*.

Généralement, la structure de la couche Grammatical Unit correspond à celle de la couche Wordin, puisqu'un Wordin est le plus souvent une unité linguistique complète. Cependant, les mots composés sont séparés en plusieurs Wordins par le pré-traitement. De ce fait, l'analyseur effectue une étape de recomposition de manière à regrouper plusieurs Wordins en une seule Grammatical Unit. La recomposition se calcule de proche en proche, en fonction de règles décrites dans un fichier. Un exemple de règle : *préfixe + tiret + infinitif* $\rightarrow$ *infinitif*. Cette régle s'applique par exemple à *sur-couver*, absent du dictionnaire mais recomposé par l'analyse.

Analyse syntaxique. L'analyse consiste, pour une suite de mots $\{w_1, w_2, \ldots, w_n\} = W$, à déterminer la meilleure suite de catégories $\{t_1, t_2, \ldots, t_n\} = T_{MAX}$. Par la règle de Bayes, ceci se formalise :

$$T_{MAX} = \arg\max_T P(W|T)P(T) \qquad (1)$$

Le modèle de langue, $P(T)$ est classiquement réduit à un modèle *n*-gramme lissé. Le lissage utilisé ici est une *interpolation linéaire*. Le modèle de mots $P(W|T)$, classiquement ignoré parce que difficile à estimer, est réintroduit sous la forme d'un modèle de classes d'ambiguïté lexicales $P(A|T)$. Nos tests, réalisés à partir du corpus d'entraînement par *10-fold-cross-validation*, montrent que le système d'analyse profite à la fois de l'interpolation linéaire sur le modèle de langue, et de la réintroduction du modèle de mots au travers d'un modèle de classes. En moyenne, le système donne 94,5% d'étiquetage grammatical correct. Un article complet est dédié à notre analyse syntaxique [2].

Conversion graphèmes-phonèmes. Le module produit la transcription phonétique des mots du texte. Pour chaque graphème d'un mot considéré hors contexte, le système décide, à l'aide d'un arbre de décision compilé à partir d'un dictionnaire d'apprentissage [9], du phonème qui lui correspond en tenant compte de son contexte graphémique et de la nature lexicale du mot.
Le mot est ensuite remis en contexte, où les phénomènes phonétiques traités sont :
– la liaison : les oiseaux $\rightarrow$ [l e z w a z o]
– l'amuïssement : cinq mille $\rightarrow$ [s ɛ̃ _ m i l]
– la lubrification du discours : quelques patients $\rightarrow$ [k ɛ l k ə p a s j ɑ̃], quelques amis $\rightarrow$ [k ɛ l k ə z a m i]

Génération de la prosodie. Le module se divise en 2 parties : prosodie linguistique et prosodie acoustique. La prosodie linguistique génère les groupes de souffle (*Rhythm Group*) et la syllabation des phrases, à l'aide d'algorithmes de type chink/chunk. Ces informations sont suffisantes pour la synthèse par sélection d'unités non-uniformes via LiONS [6], mais incomplètes pour la synthèse par concaténation de diphones via MBROLA.
La prosodie acoustique, dans eLite, est basée sur un arbre de régression et de classification (CARTs), entraîné sur un corpus de parole qui, pour chaque phonème du corpus, recense le nombre de phonèmes et l'accent symbolique de la syllabe à laquelle il appartient [8]. Pour un phonème donné, le CARTs détermine sa durée, mais également sa fréquence fondamentale s'il s'agit d'un phonème voisé.

4. SÉLECTION ET DSP

MBROLA fait partie des premiers systèmes de synthèse à utiliser une base de données vocales pour la génération du signal de parole. Dans ces systèmes, un seul exemplaire de chaque unité de parole (généralement des diphones) est représenté dans la base. L'idée sous-jacente est de regénérer les informations prosodiques (F0, durée) au moment de la synthèse dans le DSP. Malheureusement, ce traitement du signal entraîne une détérioration de la qualité et du naturel de la voix de synthèse.
Pour conférer à la synthèse un caractère plus naturel, voire proche de la parole humaine, les chercheurs ont voulu mettre en œuvre le principe de *choose the best to modify the least* [1] : au lieu de ne contenir qu'un exemplaire des unités de parole, le corpus employé en compte plusieurs, non neutralisés et donc prosodiquement variables. Cette variabilité dans la base a permis à la phase de sélection de rechercher les unités de parole qui correpondent au mieux aux unités décrites par le NLP (coût cible), et qui se concatènent au mieux (coût de concaténation) de manière à éviter autant que possible les modifications du signal. C'est cette méthode de recherche, basée sur ce double coût, qui a donné naissance à la notion d'*unités non-uniformes*. LiONS et TP-MBROLA appartiennent à ce dernier état de l'art.

MBROLA. La sélection qui précède MBROLA est minimale. Elle consiste simplement à générer un fichier dans lequel chaque ligne décrit un phonème : son nom, sa durée et, si le phonème est voisé, l'évolution de sa fréquence fondamentale.
MBROLA (*Multi-Band Re-synthesis OverLap Add*) travaille en deux phases [7]. Dans un premier temps, il extrait d'une base de données les unités acoustiques décrites dans le fichier de la sélection, et leur applique la prosodie

demandée. Dans un second temps, ces unités sont concaténées, et un lissage spectral est entrepris aux frontières des unités, de manière à éviter toute discontinuité acoustique.

LiONS et TP-MBROLA. LiONS réalise la phase de sélection des unités non-uniformes [6]. L'originalité de ce système est qu'il ne sélectionne pas les unités de parole sur la base de critères acoustiques (F0, durée, tons), mais sur la base d'informations linguistiques uniquement : position du phonème dans la syllabe, de la syllabe dans le mot et dans le groupe rythmique, contexte articulatoire, etc. Cette approche, qui donne une plus grande liberté à la courbe mélodique en évitant de répéter à l'infini les mêmes patrons prosodiques de phrase en phrase, donne à la voix de synthèse un plus grand naturel.
Le synthétiseur TP-MBROLA (*True Period MBROLA*) [4] se contente d'extraire de la base de données les unités choisies par la sélection, et les concatène en appliquant un lissage de type Overlap and Add [5] uniquement aux frontières des unités.

5. Evaluation du NLP

Le NLP a été évalué sous Windows, sur une architecture Intel Pentium Mobile 1.7 GHz pourvue d'1 Go de RAM. Les bases de données du NLP, non optimisées, représentent 3,5 Mo sur le disque et 35 Mo en RAM pour le français, 4 Mo sur le disque et 36 Mo en RAM pour l'anglais. Leur chargement en RAM prend 1,06 sec.
Les performances du NLP ont été évaluées sur un texte contenant 69.033 mots (427.426 caractères), soit environ 8 heures de parole. Sur ce corpus, le temps de traitement est de 93,684 sec, ce qui représente : 4.562,423 caractères/sec ou 736,871 mots/sec. En termes de durée de parole générée par seconde, le NLP est donc environ 306 fois temps réel.

6. Démonstrateurs

Trois démonstrations figurent sur www.multitel.be/TTS :
- *eLite Demo*. Cette application téléchargeable est disponible en français et en anglais, et fonctionne sous Windows et Linux. eLite Demo intègre uniquement MBROLA, et inclut : les bases de données nécessaires, une interface graphique de test et une bibliothèque dynamique qui peut être intégrée dans un programme tiers. Il s'agit d'une version ralentie d'eLite, qui ne peut être employée à des fins commerciales ni militaires.
- *eLite OnLine*. Il s'agit d'une interface de démonstration en ligne intégrant également MBROLA. L'utilisateur dispose ici d'une palette plus vaste de voix de synthèse (7 pour le français et 4 pour l'anglais), et peut obtenir les résultats d'analyse du texte module par module.
- *LiONS*. Des échantillons de la synthèse obtenue à partir de LiONS peuvent être écoutés sur notre site.

7. Conclusion et perspectives

eLite est un système complet de synthèse de la parole à partir du texte, qui accorde une importance certaine à l'analyse linguistique, étape-clef du processus de synthèse. Initialement dédié au synthétiseur MBROLA, eLite intègre dans sa dernière version une synthèse par sélection d'unités non-uniformes, via LiONS et TP-MBROLA.

Les perspectives d'évolution d'eLite sont importantes. Les sujets d'étude du groupe sont actuellement :
- *La correction orthographique*. Le problème en synthèse se situe au niveau des fautes audibles comme l'absence d'un accord ou une mauvaise épellation.
- *La détection de mise en page*. Un document est un ensemble de blocs : paragraphes, tables, adresses, signatures, colonnes. Détecter ces blocs permettra d'éviter une lecture linéaire et absurde du document.
- *La synthèse émotionelle*. L'idée est d'insérer de l'émotion (colère, joie, tristesse, etc.) dans la parole de synthèse obtenue par sélection d'unités non-uniformes.

8. Remerciements

Nos plus vifs remerciements vont à Vincent Pagel, qui a lancé eLite et a développé la première version du phonétiseur, et à Xavier Ricco, auteur de la première version de l'analyseur morphologique. Nous tenons à remercier tout particulièrement Fabrice Malfrère, qui a développé le générateur de prosodie acoustique, Vincent Colotte, qui a travaillé à la mise au point de LiONS, et Baris Bozkurt, auteur de TP-MBROLA. Nous remercions également Thierry Dutoit, dont les conseils avisés guident nos recherches. Enfin, nous adressons notre sincère reconnaissance à Piet Mertens, Dominique Wynsberghe et Michel Bagein, qui ont eu la patience de récolter et de construire les bases de données de l'analyseur morphologique.

Références

[1] M. Balestri, A. Pacchiotti, S. Quazza, P.L. Salza, and S. Sandri. Choose the best to modify the least : A new generation concatenative synthesis system. In *Proceedings of Eurospeech'95*, volume 1, pages 581–584, Madrid, Spain, 1999.

[2] R. Beaufort, T. Dutoit, and V. Pagel. Analyse syntaxique du français. Pondération par trigrammes lissés et classes d'ambiguïtés lexicales. In *Actes des JEP*, pages 133–136, Nancy, France, 2002.

[3] A.W. Black, P. Taylor, and R. Caley. *The Festival Speech Synthesis System : System Documentation*. University of Edinburgh, 1997.

[4] B. Bozkurt, C. d'Alessandro, T. Dutoit, V. Pagel, and R. Prudon. Improving Quality of MBROLA Synthesis for Non-Uniform Units Synthesis. In *Proceedings of the IEEE TTS 2002 Workshop*, 2002.

[5] F. Charpentier and M. Stella. Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In *Proceedings of ICASSP'86*, pages 2015–2018, Tokyo, Japan, 1986.

[6] V. Colotte and R. Beaufort. Synthèse vocale par sélection linguistiquement orientée d'unités non-uniformes : LiONS. In *Actes des JEP*, Fès, Maroc, 2004.

[7] T. Dutoit and V. Pagel. Le projet MBROLA : vers un ensemble de synthétiseurs vocaux disponibles gratuitement pour utilisation non-commerciale. In *Actes des JEP*, pages 441–444, Avignon, France, 1996.

[8] F. Malfrere, T. Dutoit, and P. Mertens. Fully Automatic Prosody Generator for Text-to-Speech Synthesis. In *Proceedings of ICSLP*, pages 1395–1398, Sydney, Australia, 1998.

[9] V. Pagel, K. Lenzo, and A. W. Black. Letter-to-Sound Rules for Accented Lexicon Compression. In *Proceedings of ICSLP*, pages 252–255, Sydney, Australia, 1998.