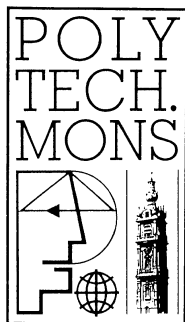


Rapport d'activité MULTITEL - TCTS

Groupe «Synthèse Vocale»



T. Dutoit

Faculté Polytechnique de Mons

MULTITEL ASBL

Période du 01-07-00 au 31-12-00

Composition du groupe TTS au 31/12/00

Personnel :

M. Bagein	ingénieur de recherche, Multitel
R. Beaufort	linguiste-informaticien de recherche, Multitel (arrivé le 1/10/2000)
B. Bozkurt	ingénieur de recherche, Multitel (arrivé le 20/08/2000)
T. Dutoit	chef de groupe, Chargé de cours FPMs
X. Ricco	ingénieur de recherche, Multitel (arrivé le 01/08/2000)
D. Wynsberghe	ingénieur de recherche, boursière FIRST-Doctorat-entreprise FPMs-Babel Technologies

Visiteurs :

A. A. Arman	professeur visiteur, Institut Teknologi Bandung- INDONESIA 15/07/00 – 15/10/00
C. Citta	stagiaire DEC2 en Ingénierie Linguistique 01/11/00 – 31/12/00
S. Rocca	stagiaire ENSERG Grenoble 01/07/00 – 30/09/00

Collaborateurs :

F. Malfrère	doctorant FPMs; employé par Babel Technologies S.A.
P. Mertens	collaborateur scientifique, KUL
O. Platteau	collaborateur bénévole
G. Casse	collaborateur bénévole

0. Introduction

Depuis 5 ans, le groupe TTS mène de front plusieurs projets très complémentaires, mis au point selon des principes similaires à ceux qui ont prévalu lors de la conception du projet MBROLA :

- Mise à disposition gratuite (pour utilisation non-commerciale, non-militaire) de certains résultats des recherches du service (*logiciels/bases de données*), dans le but d'attirer autour de ces projets de nombreux partenaires extérieurs, et construire ainsi une plate-forme logicielle solide et réellement multi-lingue pour la synthèse vocale.
- Incitation à participation extérieure, par proposition de *services* gratuits, à la condition que les résultats puissent être mis à disposition des scientifiques dans le cadre du projet. Les logiciels concernés sont ceux qui impliquent une part importante de la propriété intellectuelle ou du savoir-faire du service, ce qui en rend délicate la distribution gratuite sur internet. Les services consistent alors à utiliser les logiciels propriétaires sur des données fournies par des tiers.

Les projets que nous menons actuellement sont au nombre de 5 : MBROLA, EULER, NUMBROLA, MBRDICO, et HOOK. Nous en donnons l'état d'avancement dans les sections suivantes, et terminons par une liste de publications.

1. MBROLA

(<http://tcts.fpms.ac.be/synthesis/mbrola>)

L'algorithme de synthèse MBROLA est à présent disponible en **23 langues** et **38 voix**, grâce à diverses collaborations officielles (contrats de licences croisées signés).

Les voix récentes et en cours sont résumées sur le tableau ci-dessous :

Database Name (nv: update of an existing database)	Getting ready for recording (diphone list, text to read, etc.)	Recording, preproces sing	Segmentation	Mbrolization	Testing	Putting on web (with info files, flags, etc.)
Ar2 (Arabic)				*	*	*
Cz2 (Czech)-nv				*		*
Fr6 (French)		*				
Fr7 (French)		*				
Fr8 (French)		*				
In1 (Hindi)-nv				*		*
In2 (Hindi)						*
Id1 (Indonesian)		*		*	*	*

It1 (Italian)		*		*	*	*
It2 (Italian)		*		*	*	*
NI3 (Dutch)-nv				*		*
Nz1 (Maori)-nv				*		*
Tr1 (Turkish)	*	*	*	*	*	*

* represents the processes which are actively joined or completed alone.

List of on-process contacts: Sinhala, Greek

List of initialized contacts: Persian, Ukrainian, Finnish, Farsi, Yoruba, Tamil, Serbian, Spanish

Des améliorations de la technique de synthèse MBROLA sont actuellement à l'étude (B. Bozkurt). Un synthétiseur de parole basé sur une modélisation harmonique a été réalisé ; il utilise le résultat des analyses effectuées par les modules d'analyse de MBROLA, mais implémente la synthèse dans le domaine fréquentiel. Le but est de tester différentes hypothèses sur les améliorations à apporter à la qualité de la parole de synthèse. Une des voies explorées concerne l'étude des phases des harmoniques, pour lesquelles un outil de visualisation et d'étude a été conçu (sous Matlab). Une autre piste concerne l'étude des ondelettes et leur utilisation en synthèse vocale.

2. Euler (<http://tcts.fpms.ac.be/synthesis/euler>)

EULER est une plate-forme logicielle en synthèse de parole multilingue. EULER est donc le prolongement direct de MBROLA, qui n'en est que la partie « synthèse à partir d'une entrée phonétique », alors qu'EULER est un véritable synthétiseur du texte à la parole. Il constitue véritablement le cœur de l'activité de synthèse du laboratoire, puisque tous les autres projets en sont des clients.

Les composants d'EULER sont :

- Un **noyau** gérant plusieurs langues et locuteurs.
- La **MLC**: Multi Layer Container (C++, extension du STL).
- Un **pré-processeur (hard-coded)**.
- Un **analyseur morphologique (par règles)**.
- Un **classificateur statistique (n-gram)**
- Un **phonétiseur** générique (par règles ou corpus-based).
- Un module de **post-phonétisation (hard-coded)**.
- Un système de **génération de tons (hard-coded)**.
- Un système de **génération de prosodie (corpus-based)**.
- Un **synthétiseur** phonèmes vers parole (**MBROLA**)

Le projet EULER, par lequel le laboratoire TCTS met le logiciel EULER à disposition du grand public pour toute utilisation non commerciale et non militaire, a été officiellement lancé sur la mailing liste MBROLA (plus de 300 utilisateurs) le mercredi 26/05/99, dans sa première version.

Début novembre 2000, nous avons finalisé et rendu disponible la version 2.00 bêta d'EULER. EULER implémente désormais un synthétiseur TTS complet en français et en arabe et un synthétiseur par mots (dictionnaire parlant) en anglais américain et insulaire, allemand, espagnol, néerlandais et turc. La documentation est en cours (elle a été fortement revue dans le sens d'une meilleure convivialité). La version finale 2.00 (c.-à-d. la version actuelle avec sa documentation) est prévue pour fin mars 2001.

Nous travaillons d'ores et déjà à la mise au point de la version 2.1 d'EULER. Ainsi, un phonétiseur en Turc a été mis à point (technologie ID3 - arbres de décision -voir plus loin : projet mbrdico) en décembre 2000, et intégré à EULER. Nous avons également adjoint à EULER un interpréteur prolog (issu de SWI Prolog), qui permet d'attacher à EULER des algorithmes écrits dans ce langage (très utilisé en intelligence artificielle). Le but est de pouvoir faire du prototypage pour certaines fonctions plus linguistiques faisant appel à de l'expertise.

Parmi les tâches qui restent à réaliser si l'on veut maximiser l'utilisabilité d'EULER, il faudra considérer comme prioritaires celle qui concerne l'accès aux données de la MLC depuis un langage interprété, et la mise au point d'applicatifs externes pour chaque module d'EULER séparément(comme l'est déjà MBROLI pour MBROLA). Nous sommes

par ailleurs occupés à développer un accès JSML à EULER, pour le rendre compatible à la norme en vigueur depuis peu.

Un portage Mac est également prévu.

3. NUMBROLA

La synthèse de la parole connaît en ce moment une véritable révolution technologique, semblable à celle qu'a connue la reconnaissance il y a une dizaine d'années. L'innovation provient fondamentalement de la preuve, faite tout récemment, qu'il est possible de synthétiser de la parole tout à fait naturelle à partir d'unités vocales choisies automatiquement dans une base de données d'à peine une heure de parole, et dont on modifie la durée et/ou l'intonation. Le projet NUMBROLA porte sur les algorithmes permettant ce choix automatique, ainsi que sur les modifications à y apporter sans introduire de dégradation audible. Le but ultime est de mettre au point un logiciel de synthèse multilingue basé sur ce principe, et dont les performances seront testées pour le plus grand nombre de langues et de voix possible.

Apparue pour la première fois en 1995 (et sous une forme très embryonnaire), la synthèse par sélection s'est surtout développée en 1998 (elle a fortement impressionné lors du récent workshop international sur la synthèse, à Sydney, en novembre 1998). Des prototypes sont à l'heure actuelle étudiés pour la plupart des langues occidentales et le japonais.

Deux versions de ces algorithmes sont actuellement commercialisées : RealSpeak de Lernout & Hauspie, et Speechify issue de la technologie développée chez AT&T Shannon Labs et commercialisée à la société américaine Speechworks.

Le projet NUMBROLA, initié en juin 1999, se propose d'étudier l'utilisation de MBROLA comme base pour la manipulation des signaux dans le cadre de ces techniques de synthèse par sélection d'unités dans de grandes bases de données.

Le but est de mettre au point des synthétiseurs dans le plus grand nombre de langues/voix possibles.

Ce projet n'a pas évolué significativement depuis juin 2000, l'essentiel du temps de travail du groupe ayant été consacré à EULER 2.00.

Un ensemble de tests préliminaires ont été effectués :

1. Simulation de sélection d'unités

Méthode :

Prendre une phrase;

Constituer une base de données MBROLA sur base des phonèmes contenus dans cette phrase;

Resynthétiser cette phrase avec la nouvelle base.

Résultats :

Qualité de parole de synthèse meilleure qu'avec la concaténation de diphtongues.

2. Effet de la modélisation du pitch:

Méthode :

La qualité de la parole de synthèse par sélection d'unités est sensible aux marqueurs de pitch.

(tests de qualité de synthèse par sélection d'unités sur base d'un fichier .pho avec pitch basé sur un modèle acoustique et d'autres tests de qualité sur base d'un fichier .pho avec pitch basé sur un modèle prosodique perceptuel.)

Résultats :

Sur 7 tests (morceaux de phrases), 5 donnent une meilleure qualité de synthèse dans le cas du modèle acoustique et 2 dans le cas du modèle perceptuel. Globalement, le modèle acoustique semble donc mieux refléter la réalité.

Deux types de f0 stylisées ont été étudiés : modèle acoustique et modèle perceptuel. Les résultats ont été comparés à ceux obtenus avec des f0 brutes.

Le modèle perceptuel apparaît comme le moins bon. Suivant les phrases testées, le modèle acoustique ou les f0 brutes donnent de meilleurs résultats. *La parole de synthèse devient quasi indissociable de la parole naturelle même si le pitch n'est pas toujours conservé par rapport à l'originale* (on a donc deux exemplaires différents mais on ne sait distinguer qui est quoi).

A partir de ces tests de validation, le laboratoire a déjà développé la ressources suivantes: le système de sélection d'unités. Ce système, générique, réalise l'extraction rapide d'unités non uniformes à partir de données linguistiques et acoustiques. Ce système accepte différentes granularités d'unités : le demi-phonème, le phonème et le diphone.

Nous allons consacrer à ce projet une partie importante du temps disponible pour les prochains mois, dans deux directions :

1. Adaptation de MBROLA au formalisme NUU (adaptation des outils de mbrolisation)
2. Implémentation d'un calcul de distance entre parole de synthèse et parole naturelle. Ce problème est en effet d'une importance cruciale pour la qualité de la synthèse produite. L'implantation doit se faire sous MATLAB. Les mois à venir seront donc consacrés à la programmation des idées retenues dans la littérature spécialisée.

4. Le projet W/HOOK

(<http://tcts.fpms.ac.be/synthesis/w>)

Le projet W ("w" étant l'abréviation du mot "word" en langage braille) a pour but de tirer profit des recherches faites au laboratoire en synthèse vocale pour en constituer des applications utiles pour les personnes handicapées de la parole. Dans ce cadre, nous avons mis au point un logiciel qui permet aux personnes handicapées de bénéficier du synthétiseur vocal MBROLA pour participer en temps réel à des discussions orales. Le logiciel incorpore à cet effet une interface conviviale entre l'utilisateur et le synthétiseur. Le principal inconvénient lors de l'utilisation d'un synthétiseur vocal en temps réel est la vitesse de frappe au clavier. En effet, si une personne veut utiliser un synthétiseur vocal pour participer à des discussions orales, il faut que la vitesse de frappe d'un mot au clavier avoisine la vitesse d'élocution de ce mot. Pour augmenter cette vitesse de frappe, W utilise des abréviations des mots les plus couramment utilisés dans le langage oral. Ainsi, la personne handicapée doit uniquement taper l'abréviation du mot désiré pour en produire la synthèse vocale.

Pour éviter de créer une liste d'abréviations propre au logiciel, il a été décidé d'utiliser les contractions du **Braille grade II**. Ce langage est utilisé depuis des décennies par les personnes aveugles pour lire et écrire rapidement les textes au clavier. Il a donc déjà fait ses preuves en ce qui concerne la réduction du temps de saisie d'un texte. Il est clair que toutes les lettres en Braille ont leur équivalent en lettres standards, ce qui a permis de dresser une liste de contractions de mots (comme "boulevard" abrégé en "bd") et de groupe de lettres standards (comme "ment" abrégé en "m").

W est disponible gratuitement (sources et exécutable; licence GNU) sur le serveur de la FPMS, à l'adresse <http://tcts.fpms.ac.be/synthesis/w>.

Le logiciel W a été intégré au noyau de Windows NT. Le logiciel capture désormais directement toute frappe au clavier, et propose automatiquement dans une fenêtre pop-up la transcription la plus probable au vu de ce qui a déjà été frappé au clavier. Ce nouveau logiciel, qui n'inclut plus directement de synthèse vocale, mais peut être utilisé en combinaison avec EULER, porte le nom de HOOK.

Le travail que nous avons mené sur W et HOOK ces dernières années nous a permis d'entrer dans un consortium européen pour la mise au point du projet FASTY, qui commencera le 1 janvier 2001 et pour lequel nous travaillerons à raison de 32 hommes.mois.

Avancement récent sur le projet HOOK

- Hook est maintenant dans la traybar de Windows.
- Il fonctionne en tâche de fond.
- Le programme hook.exe (Interface utilisateur) exécute hooksrv.exe.

- Hooksrv.exe est un serveur contenant le noyau de hook. Il effectue des traitements à l'aide d'un système de résolution par règles multi-niveaux (MLRR).
- Il ne s'agit pas d'« autohook » plus.

Prévisions

- Gestion personnalisée de bases de données dans hook ...
L'utilisateur pourra gérer diverses bases de données à tout moment en navigant dans le menu de hook (systray) et en choisissant une base personnelle ou la base de phonétisation ou encore une base de désabréviation (braille grade II).
- Hook dépendant du clavier de l'utilisateur, il serait intéressant de pouvoir gérer le type de pays associé au clavier.
- Documentation de Hook.
- Site web de Hook.
- Package de Hook.

5. MBRDICO

(<http://tcts.fpms.ac.be/synthesis/mbrdico>)

En collaboration avec Kevin Lenzo (Carnegie Melon) et Alan Black (CSTR/Edimbourg) nous avons développé en 1998 une méthode qui dérive automatiquement les règles de transcriptions, encodées dans des arbres de décision.

Ce système appelé MBRDICO a été utilisé pour traiter l'anglais américain, l'anglais britannique, l'arabe, l'espagnol, le français et le néerlandais.

Trois nouvelles ressources ont été apportées à ce projet :

- Une révision complète du corpus français a été réalisée.
- Une version d'épellation est réalisée.
- Une version en Turc vient d'être réalisée.

L'ensemble des logiciels (sources et exécutables sous Win95/NT et Sun/Solaris) et des résultats a été mis à disposition sur internet, sous licence GNU.

6. Publications

Articles dans les actes de conférences (voir annexe1 pour les compte-rendus de ces conférences)

M. BAGEIN, T. DUTOIT, F. MALFRERE, V. PAGEL, A. RUELLE, N. TOUNSI, D. WYNSBERGHE, 2000, , "The EULER Project : an Open, Generic, Multi-lingual and Multi-Platform Text-To-Speech System" , Proc. ProRISC'2000, Veldhoven, December 2000, pp.193-197.

5th Year MBROLA-CDROM

Nous avons compilé l'ensemble des projets MBROLA et EULER sous la forme d'un CDROM de vœux (pour l'année nouvelle 2001). Ce CDROM chantant (une chorale de synthèse MBROLA a été spécialement créée pour l'occasion ; voir <http://tcts.fpms.ac.be/synthesis/mbrola/xmas.mp3>) a été envoyé à plus de 100 personnes, liées de près ou de loin au projet MBROLA, ainsi qu'aux contacts du centre de recherche.

Annexe 1 : compte-rendus de conférences et expositions

ProRISC'2000, Veldhoven, December 2000 (M. Bagein, T. Dutoit)

Cette conférence d'une journée rassemblait des chercheurs et doctorants dans le domaine général du traitement du signal. Une session y était consacrée aux outils logiciels récemment développés. Nous y avons présenté le logiciel EULER. Nous avons également assisté à la présentation remarquable du Prof. B. Macq (UCL), sur le watermarking numérique.

SpeechTek'2000, New York, 1-3 November 2000 (T. Dutoit)

SpeechTek is one of the major yearly commercial events in speech technology. All companies with some activity on speech processing exhibit there, and it is a good occasion to take the temperature of the speech market.

Here is a list of hot spots at this year's exhibition:

-Voxi, (www.voxi.com) : Swedish company, proposes a tool for building dialogue systems from concepts.

-VerbalTel (www.verbaltek.com), American Company, proposes speech recognition on PDAs and wireless phones.

-Locus (www.locusdialogue.com), Canadian company, sells its own speech rec and BABEL's speech synthesis to canadian and american clients (among which Bell Atlantic!). I was impressed by the seriousness of this company. They had a big booth at spechtek, and they are an ideal connection between the French speaking and American speaking markets.

-Wavemakers (www.wavemakers.com), Canadian company, sells a patented speech de-noising technology. They have an impressive demo : they send the same wave file to 2 computers at a time, both running the L&H recognizer, but one of the inputs is previously passed through their system. The resulting difference is impressive.

-Forcecomputers (www.forcecomputers.com). This company has been created by the initial developers of dectalk (rule-based synth), previously sold by Compaq, and abandoned some years ago. It now comes back in force, with extension to new languages via (among other things) university collaborations.

-Lipsinc (www.lipsinc.com), American company, offers high quality virtual talking heads, with lips synchronized with natural speech pronounced by a human. Very impressive.

-L&H, (www.lhsl.com) was presenting its realspeak TTS (available in French, German, US English, other languages under development). This clearly attracted many people.

-Speechworks (www.speechworks.com), American company, offers the AT&T TTS in US English. This is clearly the very best TTS commercially available. Speechworks had a big booth at spechtek (though less big then L&H's).

-Philips : they had one of the biggest booths, many people on the stand, but showed little things, and mostly talked among themselves...

-Babel. The Babel booth attracted many people (it was also nicely situated, just close to L&H, Speechworks, and Philips). Babel gave demos of mbrola-based TTS, Talking heads (impressive video from BBC4) and speech rec (interactive kiosk).

Plus some companies selling head-mounted mics (www.emkayproducts.com , www.andraelectronics.com , www.gnnetcom.com)

Some comments on hot topics :

PDAs : I saw two companies with a clear target to PDAs : VerbalTek (speech rec) and Force (speech synth).

In the US, the PalmVII is out now, and seems to sell very well. The PALM VII has a built in wireless connection. Other PDAs are announced with the same feature. This

suddenly opens new markets, as wireless phones and PDAs are clearly doomed to become one single tool. PDAs already offer virtual on-screen phone keyboards, for dialing wireless... DecTalk claims it is one of the only synths to be ported to PDAs, because of its compact size (rule-based). MBROLA clearly also has a bright future on PDAs, given its compression capabilities. Babel is currently working on integrating it on such devices.

Visit at SpeechWorks, NYC office, 3 November 2000 (T. Dutoit)

I met Roberto Pieraccini at SpeechTek, a former colleague of mine when we were at AT&T, who now heads the dialogue research group at SpeechWorks. He told me that the serious competitors of SpeechWorks are Nuance in the US, and L&H+Phillips in Europe. Roberto invited me to visit the NYC offices of SpeechWorks (which I did on Friday, 4th), and we discussed about the company. SpeechWorks is now expanding to European offices (Paris soon, and one is already open in Germany). It employs about 500 people, among which MORE THAN 60% are commercials (!). Speechworks has acquired a NON EXCLUSIVE license of AT&T's TTS technology by giving AT&T some shares of the company.

Visit at AT&T Shannon Labs, Florham Park, NJ, 7 November 2000 (T. Dutoit)

AT&T has just split into 4 companies. The research labs seem not to have been really taken into account in the split. It is now called AT&T Labs Co., and has been given 3 years to become self-supported, by selling licenses of patents.

In the morning, I attended a talk by ... HERVE BOURLARD (!), on recent developments in Speech Rec at Idiap. This talk was very much appreciated at AT&T, for its prospective view. See below for a summary.

I also gave a talk in the afternoon, on Speech Synthesis at TCTS-MULTITEL.

I explained the history of MULTITEL, and its expected future.

I then spent 1h30 to explain the current state of our main projects : MBROLA, EULER, MBRDICO, and W.

One thing I understood from discussions is that the quality of AT&T's TTS RELIES MORE AND MORE ON THE (INCREASING) SIZE AND ADEQUACY OF THE SPEECH DATABASE TO THE USE OF THE SYNTHESIZER.

HB's Talk

=====

Herve talked about two major works at IDIAP :

1. MultiStream. Hr showed a theoretical, statistical approach to deriving a multiband speech recognizer which naturally produces error rates as the product of error rates from single-band recognizers.

This is achieved by establishing the partition of all possible recognizers using 0 band, 1 band, 2 bands, etc., and adequately using the probabilities they produce. He showed that this led to a global error rate which is the product of all the error rate of each individual speech recognizer, which is known empirically as the best way to combine recognizers (this comes from Fletcher's work in 1929!).

2. HMM2. Herve has presented a great idea on how to find dynamic bands or streams (rather than using static, fixed bands), by using a "frequency" HMM on the spectral coefficients of each frame.

This HMM actually automatically performs the segmentation of the spectral coefficients into bands, which can then be used multiband speech recognition. Herve showed that his HMM2 model actually led to a segmentation into formant bands... very interesting.