Activities

MULTITEL is a research centre that stepped out of the Faculté Polytechnique de Mons (FPMs) in 1995. MULTITEL maintains a strong collaboration with one of its founder laboratories, namely the "Théorie des Circuits et Traitement du Signal" laboratory of FPMs, which has been active in speech processing and more especially in Text-to-Speech synthesis (TTS) from the mid of ‘90. MULTITEL has been involved in several intern and national research projects (cf. table I) and it has gained great expertise in Text-to-Speech synthesis. In 2000, MULTITEL became an independent nonprofit research centre. Nowadays, the activities of MULTITEL in Text-to-Speech synthesis are focused on applied research, software development and consulting.

Applied Research

This activity consists in studying linguistic theories, computer-based and algorithmic principles, and mathematical bases and concepts required for developing TTS software. Apart from a technological survey activity, MULTITEL is currently interested in Non-Uniform Units (NUU) based synthesis, in spell-checking, and in automatic text-pattern recognition (sender, addressee, address, table columns,…).

Software Development

Research at MULTITEL in Text-to-Speech synthesis from a few years has led to the development of eLite (cf. table II). Based on the OOBP (Object Oriented Block Programming) paradigm, the eLite software is a complete Text-to-Speech synthesis system, and provides a useful platform for the development and the assessment of new algorithms used in synthesis. eLite includes Eliot (cf. table III) for the Natural Language Processing (NLP), whereas the Digital Signal Processing (DSP) is performed by either MBROLA, or the combination of a non-uniform units selector, LIONS, and a MBROLA-derived algorithm, TP-MBROLA.

Consulting

Besides research and development activities, MULTITEL provides consulting in speech synthesis: help for software development (specification and prototyping), textual data collection for new languages designing, etc.

### Table I: ongoing and past TCTS – MULTITEL projects involving Text-to-Speech synthesis.

<table>
<thead>
<tr>
<th>Project</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYPOLE RW</td>
<td>Système POrtable pour LEcture de documents (portable systems for documents reading)</td>
</tr>
<tr>
<td>LIONS projet interne</td>
<td>Linguistically-Oriented Non-uniform-units Selector</td>
</tr>
<tr>
<td>MODIVOC RW – INITIATIVE</td>
<td>systèmes MOobiles et Distribués à Interface VOCale (mobile, distributed systems with vocal interface)</td>
</tr>
<tr>
<td>eLite projet interne</td>
<td>Enhanced, Linguistically-based TExt-tospeech synthesizer</td>
</tr>
<tr>
<td>Eliot intern project</td>
<td>Electronic processing Linguistically-Oriented of Texts</td>
</tr>
<tr>
<td>MBROLA intern project</td>
<td>Multi Band Resynthesis with OverLap Add</td>
</tr>
<tr>
<td>MLRR intern project</td>
<td>Multi Layers Rewriting Rules</td>
</tr>
<tr>
<td>W intern project</td>
<td>the Word project</td>
</tr>
<tr>
<td>EULER intern project</td>
<td>unified, free, extensible research environment for Text-to-Speech synthesis</td>
</tr>
<tr>
<td>MBRDICO intern project</td>
<td>talking dictionary using MBROLA as a back-end synthesizer</td>
</tr>
<tr>
<td>MBROALIGN intern project</td>
<td>fast MBROLA-based Text-to-Speech aligner</td>
</tr>
</tbody>
</table>

### Table II: Technical features of eLite.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLP</td>
<td>Eliot</td>
</tr>
<tr>
<td>DSP</td>
<td>MBROLA</td>
</tr>
<tr>
<td>Plateorm</td>
<td>Windows 98, 2000, NT4.0, XP Linux MacOS X</td>
</tr>
</tbody>
</table>

RW : Région Wallonne
### Preprocessing

- Regular expressions-based linguistic units detection. Regular expressions are compiled in a unique Finite State Automata

### Morphological Analysis

- Word lemmatization
- Analysis of complex units subcomponents (URLs, numbers, …)
- Diacritic accents restoration

### Syntactic Analysis

- Language model: tri-gram with linear interpolation
- Word model: lexical ambiguity classes
- Viterbi decoding for words in context

### Grapheme-to-Phoneme Conversion

- Induction Decision Tree (ID3) combined with an exceptions dictionary
- Unpronounceable words spelling
- Liaisons and discourse fluidity management

### Prosody Generation

- Acoustic features generation (fundamental frequency, duration), using corpus-based Classification And Regression Trees (CARTs) training
- Linguistic prosody

**Table III**: Technical features of Eliot.

---

### Demonstrations

#### Win eLite

This demonstration interface (cf. figure I), which uses eLite as a back-end synthesizer, permits the user to test the current speech synthesis state-of-the-art of MULTITEL. The main part of the interface is an edit zone where the user may both type a text or open an existing text file. Via the interface, the user can listen speech signal generated from his text. Using the menu, it is easy to act on several synthesis features: language (French or English), voice (male or female), elocution speed, pitch and volume. Using this interface, the virtual agent of eLite, called eFriend, can moreover be tested; eFriend's lips are articulated as fast as a sound is pronounced.

**Figure I**: The Win eLite interface.

### eFriend

Virtual agents of eLite (cf. figure II) have been developed to attest the fact that synchronization of the MBROLA speech signal with another application can be made in a simple way. The principle is really easy: the front-end program can ask MBROLA what phoneme is being synthesized. eFriend is a front-end program, and takes the MBROLA feedback into account to modify and adapt virtual agent's lips.

**Figure II**: eFriends of eLite, Anne-Carole and Vincent.

### SpeechKiosk

The interactive booth SpeechKiosk (cf. figure III) uses the TTS techniques that MULTITEL has developed over the last years in order to build a virtual receptionist. For example, SpeechKiosk can be located at a building entrance and the visitor can come to the booth in order to query some information or get into contact with a staff member. The visitor has just to stand in front of the booth and interact with eFriend, who answers his query using a synthetic voice. This application demonstrates a good example of a multimodal human-machine interface that is based on speech and haptic modalities (e.g., touch screen).

**Figure III**: SpeechKiosk interactive booth. The software is developed by MULTITEL and the hardware is designed by IMMEDIA.
Technology

Speech is probably the most efficient medium for communication between humans. Speech is now becoming a very efficient medium for human-machine communication too (cf. figure IV). The speaker converts the information he or she wants to transmit to the machine into speech, i.e. a sequence of structured sounds forming words and sentences. This information, extracted from the signal by an Automatic Speech Recognition system (ASR), can then be interpreted by the machine and the appropriated action may be launched. For instance, a speech message can be produced with a TTS system and emitted towards the speaker with a loudspeaker.

Modern TTS systems have a pipe-line block architecture (cf. figure V) able to read any text aloud, whether it was directly introduced in the computer by an user or scanned and submitted to an optical character recognition system [1].

Pre-processing

The goal of the pre-processor is to tokenize input text into words, symbols or group of strongly connected words, and to get rid of unwanted characters (spaces, end of line characters, special characters...). By using regular expressions, the preprocessor detects Wordsins (“everything between spaces”, tokens), but also special expressions like URLs, dates, phone numbers or punctuations. As soon as an expression is detected, the pre-processor runs a specialized expression parser with the view to find out the different units the expression contains (for instance, an URL is composed of symbols, punctuations, words, acronyms...).

Morphological analysis

At MULTITEL, morphological analysis is linguistic unit-based. A linguistic unit can involve more than one word, as for compound words, phone numbers or URL addresses. Thus, this notion implies a double morphological analysis. The first one tries to find out the list of syntactic categories attributable to each linguistic unit. The second one analyses each linguistic unit constituent and proposes the possible corresponding set of lexical parts-of-speech. The advantage of using double-level analysis is the possibility for each module to process only the information it needs. For example, the syntactic analyser will be interested in the categories, while the phonetizer will use the parts-of-speech.

Syntactic analysis

This module has to decide which category does effectively correspond to each linguistic unit, given the context of the sentence. This means the analyser has to find out the best path in a lattice of categories. To achieve this task, our statistic-based analyser combines a language model with a word model. The language model is a classic tri-gram, with the particularity to linearly interpolate the information from lower-level n-grams. Unlike most systems, our analyser keeps a word model, thanks to the concept of ambiguity classes [2]. An ambiguity class groups together all words (from the training corpus) sharing the same categories; so, the class permits to estimate the probability of each category for a word, even if this word has never been seen in the training corpus. The probability of a sentence is computed by dynamic programming, using the Viterbi algorithm.
Graphemes-to-phonemes transcription

Here, the goal is to provide the phonemic transcription for each word of the text. First, this transcription is performed without taking the word context into account. The tool is a decision tree, compiled from a training dictionary [3]. For a given grapheme, the tree chooses the corresponding phoneme by taking both graphemic contexts and lexical part-of-speech into account. Then, whether the transcription is not correct for the language, the word is spelt. Second, the word transcription is put back in its context. The pronunciation of a word can effectively be influenced by the words around. Then, those words have to be checked so that the transcription can be modified. Some frequent phenomena are liaison (les_oiseaux), deletion (dix/mille) and the discourse fluidity (compare the last “e” of quelque pronounced in “quelque_personne” but not in “quelque/ ami”).

Where a choice has to be made

After the graphemes-to-phonemes transcription step, the system can take two directions. Either the concatenation-based synthesis, that is MBROLA at MULTITEL; in this case, one needs an acoustic prosody generation module. Either the Non-Uniform Units (NUU) selection-based synthesis; in this case, a linguistic prosody is computed, then the units selection is performed, and the speech is generated thanks to the TP-MBROLA algorithm.

Acoustic prosody

A phonetic transcription is not enough. To be able to pronounce a sentence, the system needs to know the melody, the prosody to be applied. The prosodic module has to find out two features for each phoneme: its fundamental frequency and its duration. At MULTITEL, this task is corpus-based [4]: a lot of examples help to define the average frequency and duration of each phoneme given some features: the phoneme itself, the number of phonemes belonging to the same syllable, and a symbolic stress chosen with the help of Classification And Regression Trees (CARTs), also corpus-based.

MBROLA synthesis

Created in the mid of ‘90, the Multi-Band Resynthesis Overlap Add algorithm works in two steps. First, the previously computed prosody is applied to acoustic units, selected in a database according to the phonemes list to synthesize. Second, those units are concatenated and a spectral smoothing is applied to their boundaries, in order to erase discontinuities. The result is a high-quality synthetic voice, but not yet as natural as human voice.

Linguistic prosody, NUU selection and TPMBROLA synthesis

Patents are currently in the work concerning these algorithms, developed at MULTITEL. That is the reason why we only give hereunder the main principle that leads our research. The governing idea is to generate non-uniform units-based speech synthesis without acoustic information. The goal is to let the data “free”, in order to get a really variable prosody. Thus, the prosodic information used by this module is only linguistic. TP-MBROLA processes then the concatenated signal at the boundaries of the chosen units.

Applications

There exist many applications of Text-to-Speech synthesis. These applications make use of speech in order to replace or complement an interface for helping the machine to transmit an information to a person. For instance, in order to guide towards a given service or to inform that a given service is started. Sometimes, speech is the only possible interface, for example, when the speaker cannot look at the screen to read the message. Table III gives a few examples of TTS applications that demonstrate how useful text-to-Speech synthesis can be.

<table>
<thead>
<tr>
<th>Field</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Téléphony</td>
<td>Automation of telephone transactions (e.g., banking operations), automatic call centre for information services (e.g., access to weather reports), etc.</td>
</tr>
<tr>
<td>Automotive</td>
<td>Information released by in-car equipments such as the radio, the air conditioning system, the navigation system, the mobile phone (e.g., voice dialling), embedded telematic systems, etc.</td>
</tr>
<tr>
<td>Multimedia</td>
<td>Reading of electronic documents (web pages, emails, ills) or scanned pages (output of an Optical Character Recognition system).</td>
</tr>
<tr>
<td>Medical</td>
<td>Disabled people assistance: personal computer handling, domotic, mail reading…</td>
</tr>
<tr>
<td>Industrial</td>
<td>Voice-based management of control tools, by drawing operator’s attention on important events divided among several screens,…</td>
</tr>
</tbody>
</table>

Table IV : Examples of applications for Text-to-Speech synthesis

References