EULER
An Open, Generic, Multi-lingual and Multi-platform Text-to-Speech System

Thierry Dutoit, Michel Bagein, Fabrice Malfrère, Vincent Pagel, Alain Ruelle, Nawfal Tounsi and Dominique Wynsberghe

Faculté Polytechnique de Mons, Circuits Theory and Signal Processing Lab
Bâtiment Multitel, Parc Initialis, av. Copernic, B7000 Mons, BELGIUM
Tel: +32 65 374733 Fax: +32 65 374729
Web: http://tcts.fpms.ac.be
Email: {dutoit,bagein,pagel,ruelle,tounsi}@tcts.fpms.ac.be

Abstract

The aim of the collaborative project presented in this paper is to obtain a set of highly modular Text-To-Speech synthesizers for as many voices, languages and dialects as possible, free for use in non-commercial and non-military applications. This project is an extension of the MBROLA project: MBROLA is a speech synthesizer, freely distributed for non-commercial purposes, which uses diphone databases provided by users (19 languages in year 2000). Euler extends this idea to whole TTS systems by providing a backbone structure (*MLC*) and several generic algorithms for POS tagging, grapheme-to-phoneme conversion, and prosody generation. To demonstrate the potentials of the architecture and draw developpers' interest we provide a full EULER-based TTS in French and in Arabic. Euler currently runs on Windows and Linux, and it is an open project: many of its components (and certainly its kernel) are provided as GNU C++ sources. It also incorporates, as much as possible, components and data derived from other TTS-related projects.

1. Introduction

Text-to-speech (TTS) synthesis involves the computation of a speech signal from input text. TTS systems simultaneously require language processing (pre-processing, morphological analysis, part-of-speech tagging, prosodic-syntactic grouping, phonetization), acoustic-linguistic processing (for deriving phoneme durations and intonation curves) and a final digital signal processing step (speech synthesis). This makes the design of a TTS system a software engineering problem by itself.

Private and public research laboratories (from universities to telecommunication operators) have invested considerable resources in trying to design multilingual synthesizers. In most cases, however, this non-coordinated research effort has led to unavoidable cross-system incompatibility: due to an obvious lack of unified, extensible, widely-accepted and publicly available tools and databases for TTS system development, each and every synthesizer is a laboratory-specific implementation of very similar basic principles. This, in turn, has resulted in cross-language incompatibility : most multilingual TTS systems are merely collections of monolingual ones independently developed in native research labs. Not only this situation has had a negative impact on the extensibility of available TTS systems to new languages, dialects, accents, voices, and speaking styles, but it also has hampered their integration into real products: instead of incrementally refining a common, general-purpose synthesizer and providing it with high-quality interfaces for real-world applications (for handling complex text documents, for instance), research labs waste time re-implementing the wheel. Last but not least, the lack of a common backbone for TTS systems has made it very difficult to compare their quality on a module-by-module basis, thereby strongly restraining the spreading of improvements.

In contrast with this situation, state-of-the-art tools and databases for multilingual TTS system development have been recently and independently made freely available by several public research labs, as for instance:

- The Faculté Polytechnique de Mons (FPMs, Belgium) has taken an important step towards developing high-quality, multi-lingual phonetics-to-speech synthesi–zers, in the form of the MBROLA Project (Dutoit et al., 1996)[1]. The aim of this internet project is to foster international collaborations so as to obtain a set of MBROLA speech synthesizers for as many languages (including dialects) and voices as possible, free for use in non-commercial applications. 19 languages are now available with 28 voices and many more will still follow. A prosody transplantation and speech segmentation tool called MBROLIGN (Malfrère and Dutoit, 1997)[2] has also been developed and freely distributed.
- The University of Edinburgh (UED, UK) has made a major contribution to the development of open source high quality TTS. Their GPL licensed FESTIVAL (Black et al., 1997)[3] Speech Synthesis system, indeed, is nothing less than a *generic*, *modular*, *portable* and *extensible* TTS system which lays the foundations of truly global TTS research and development.
- The University of Provence (UP, France) has coordinated the MULTEXT (Véronis et al., 1994)[4] series of projects, the aim of which is to develop tools, corpora, and linguistic resources for a wide variety of languages free for non-commercial, non-military purposes.

[1] http://tcts.fpms.ac.be/synthesis/
[2] http://tcts.fpms.ac.be/synthesis/mbrolign
[3] http://www.cstr.ed.ac.uk/projects/festival.html
[4] http://www.lpl.univ-aix.fr/valorisation

The objective of EULER is to combine tools and data obtained from such major contributions into a single integrated development environment, available for Windows, UNIX and MacOS. Our hope is that the resulting research, development and production environment will reverse the disadvantageous situation outlined at the beginning of this section by cutting down the development and production prices for multi-lingual products.

2. Multi-Lingual Speech Synthesis

Euler was designed with special attention to the following three main points(Dutoit, 1996; van Santen et al., 1997):

- The quality of its phoneme-to-speech synthesis module, responsible for the segmental quality of synthetic speech. Knowing that speech is often produced by concatenating elementary speech units called segments, segmental quality in turn depends on several factors such as: type of segments chosen, speech signal model, prosody modification efficiency (strongly related to the speech model), capabilities of the segment concatenation algorithm.
- The quality of its text processing module. Text processing in a TTS system typically comprises, from beginning to end: preprocessing (correctly handle numbers, abbreviations, acronyms, etc.), morphological analysis and/or part-of-speech tagging, phonetization, syntactic-prosodic grouping, stress assignment, prosody tagging, and finally intonation and duration generation.
- The quality of its software engineering. Being very large software applications, including various sources of knowledge into a single code, and thereby typically requiring the collaborative work of specialists in many areas, the lifetime of TTS systems tends to be very sensitive to their implementation methodology and internal representation formalism. In that respect, given the various connections between the many levels of description of a language, and since the way they can be related to each other is seldom known in advance, it is common practice to organize the text and speech data handled by a TTS system in terms of multilevel data structure (MLDS), in which each level appears as an independent description of the sentence, synchronized with the other ones.

3. Multi-Level Container

The backbone of the Euler project is a C++ generic data structure called **Multi Layer Container (MLC)**, which recalls the MLDS of Festival (Black et al., 1997) and Speech Maker (van Leeuwen and te Lindert, 1993). The MLC is an extension of the C++ Standard Template Library to multi-level data with links across levels. We distribute its code under the GPL license, which makes it possible for developers to use it in their applications and to plug new modules on it.

Advantages of the MLC over a linear organization of information are numerous:

- It naturally increases readability, regarding both data and rules.
- Extensibility is greatly enhanced. Since MLCs also intrinsically admit unspecification, one can always specify additional layers to account for new analysis modules, in a way that remains transparent to previously developed modules.
- Debugging is made easier, since information provided by distinct modules is stored at different levels.
- Linear data structures make it hard to exploit additional linguistic knowledge that might be available at the input, as it is the case when speech synthesis is performed from machine-generated concepts (like in dialogue systems) rather than from plain text. TTS systems based on MLCs offer natural interfaces in all cases: synthesizing speech from concepts simply implies that the MLC is initially filled with information on other levels than the graphemic one.
- Since data is made independent of rule formalisms, cross-language portability is much better ensured.

In order to ensure maximum genericity at the module level, Euler makes use of *Dynamic Linked Libraries* on Windows and *Shared Objects* on Unix platforms. Modules are compiled libraries using a template interface to both the EULER kernel and the *MLC*. Hence, for the Euler kernel, running a TTS application merely reduces to calling a sequence of modules, defined in an initialization script. It is thus easy to build new TTS systems with different processing sequence, to replace any module in the call sequence by a new one for a performance comparison or to initialize a module with new parameters (for specifying lexicons, Ngram probabilities, diphone databases . . .).

EULER provides a library of standard algorithms (called *Engines* in our terminology) which includes at the moment an NGRAM decoder, a Multi Level Rewriting Rules interpreter (MLRR, GPL license) and a tool for building and using decision trees (ID3, GPL license). A developer can use any of those engines in his/her own modules.

The entire Euler project can be downloaded at http://tcts.fpms.ac.be/synthesis/euler, the package is auto-installable and comes with a complete user's and programmer's guide.

4. French and Arabic TTS demo

To make our point we distribute a full text to speech application in *French* and in *Arabic*, based on the EULER kernel. To demonstrate the possibilities of MLCs we also distribute a Karaoke singer (the Mbrola speech synthesi–zer was not designed to sing but it makes the point anyway) which shows how one can easily input additional layers in the MLC while running the regular TTS. This makes it possible for programmers to deal, for example, with information dedicated to facial animation. The list of modules (and related languages) that are currently available in the EULER project is given table 1.

Task	Modules [languages]
Pre-processing	RulePreprocessorFr [fr,be,ch] RulePreprocessorAr [ar]
Lexical access and Morphological analysis	RuleLemmatizer [fr]
Part-of-speech tagging	NgramTagger [fr]
Phonetizater	Phonetizer [fr,ar,en,us,es,nl] (ID3 or MLRR engine)
Prosody Generation	FMProsodyGenerator [fr,ar,es]
Phonetics to speech	MBROLA [ar, br, bz, cr, cz, de, ee, en, es, fr, gr, hb, hn, jp, mx, nl, ro, sw, us]

Table 1: Available modules and related languages

4.1. Preprocessor

A GPL licensed preprocessor is available for French, and for Arabic. It detects sentence ends, abbreviations, expands numbers and in-lexicon abbreviations. The Arabic preprocessor uses the *Multi Level Rewriting Rule* engine.

4.2. French lemmatizer

The lemmatizer provided for French is based on MORLEX, a French lexical database developed by the team of Piet Mertens (KUL, Belgium). It contains 33,000 lemmas and 160 inflexion rules, which cover about 400,000 French word forms. The lemmatization process is implemented as a reusable engine for other languages.

4.3. Ngram tagger

To develop our French part-of-speech tri-gram training corpus, we started from a 50,000 words corpus containing 4,300 sentences. It was first automatically tagged by the VERTEX chart parser for unification grammars developed by Piet Mertens. It was then manually corrected, and part-of-speecg (POS) trigrams were extracted with DARPA tools (Clarkson and Rosenfeld, 1997). On a manual evaluation for 2,000 words, an accuracy of 82% was achieved (typical errors in this experiment are reported table 2). As one can see the weakest link is the lemmatizer, since the major cause of error is an untagged word (leaving this category of errors apart, accuracy rises to 90%). Our NGRAM decoder is also a reusable engine.

Error type	Number
unknown word	171 / 2000 (8.5%)
determiner tagged as pro-clitic	74 / 2000 (3.6%)
verb tagged as noun	31 / 2000 (1.5%)
adjective tagged as noun	31 / 2000 (1.5%)
adverb tagged as noun	31 / 2000 (1.5%)
determiner tagged as noun	21 / 2000 (1%)

Table 2: French trigram-based POS tagging error rate

4.4. Phonetizer

French and Arabic text-to-phoneme conversion is achieved with our trainable ID3 algorithm, which has already been made available in the MBRDICO[5] project(Pagel et al., 1998). For French, our training corpus comprises 200,000 word forms (French-TCTS corpus) and associated morphological analysis, as we need it to disambiguate heterophone homographs. We made an evaluation on 1,000 out-of-lexicon substantives and verbs extracted from a French newspaper: the system transcribes 91% of the words according to the manual transcription. In an audio evaluation session 95% of the transcriptions were found to be acceptable (elisions and geminates explain the extra 4% in audio evaluation). Besides being a trainable grapheme-to-phoneme transcription method, this is also a good method for lexicon compression (ratio 22/1 on our French-TCTS corpus).

In French a post-phonetization module deals with *liaisons*, *élisions* and *dénasalisation* with a set of regular rules. For languages where regular grapheme-to-phoneme transcription rules can be easily designed by experts one can also use the MLRR engine (an example is provided for Arabic).

4.5. Prosody generation

Recent developments in prosody generation have highlighted the potential interest of machine learning techniques, such as multi-layer perceptrons (Traber, 1995), Classification and Regression Trees, CARTs (Hirschberg and Prieto, 1994), or other statistical techniques. A common feature of all these approaches is their training stage which requires large prosodically labeled corpora.

The approach used in EULER is based on CARTs and Non Uniform Prosodic Unit concatenation(Malfrère et al., 1998). Training is achieved with speech corpora which are automatically labeled with MBROLIGN(Deroo et al., 1998). A French speech corpus of more than one hour of read newspaper articles has been labeled and manually corrected with MBROLIGN in about 40 hours, our 40 minutes Arabic corpus required 24 hours.

The prosody generation system has been designed to be independent of the linguistic model of intonation and multilingual and has been applied for French and Arabic, the result of which can be evaluated by listening to Euler. The system is composed of four sequential stages:

- pause generation (using decision trees),

[5] http://tcts.fpms.ac.be/synthesis/mbrdico

- *prosodic phrasing module*: based on a crude chinks 'n chunks algorithm using the detection of function words for French. In Arabic, chinks'n chunks are constructed on function words, geminate consonants and long vowels.

- *F0 generation* is achieved by the concatenation of prosodic patterns automatically derived from the corpus. During the creation of the prosodic dictionary, patterns are determined by the same chinks 'n chunks algorithm as in the synthesis process. Each entry of the database is composed of prosodic marks defined on a syllable level and the intonation of each syllable is described with a set of pitch marks. During prosody generation, the choice of the sequence of patterns is based on a dynamic programming algorithm which tries to minimize F0 discontinuities at the prosodic concatenation point (final syllable of the prosodic groups), while maximizing the match between the prosodic marks to realize and those available in the pattern database.

- *duration* is derived thanks to a classification tree trained with WAGON (Black et al., 1997). Eight classification features have been chosen: the current phoneme, its class, its position in the syllable, the size of the syllable, the type of the syllable, the accent type, the position of the last accent realized and finally the phonetic class of the following phoneme.

4.6. Speech synthesis

Acoustic signal is generated by diphone concatenation thanks to the MBROLA(Dutoit et al., 1996) algorithm. A GPL module has been developed to plug MBROLA into EULER for accessing the MLC phonetic layers (phonemes and prosody).

5. Conclusions and Perspectives

EULER is intended to be a multi-lingual and multi-platform TTS publicly available for non-commercial use. It has been designed for both TTS users and TTS developers:

- Its GNU kernel lets TTS users define TTS systems in a purely declarative and unified way, while hiding their internals.

- Its MLC provides unified and open data exchange between TTS modules.

- Its generic engines are aimed at easing the extension to other languages.

What we target in the years coming is a unified Windows / MacOs / Unix TTS development environment, compatible with Festival file formats, using MBROLA as a multi-lingual phonetics-to-speech synthesizer, and readily useable in a (maximally large) number of languages made available by EULER partners (French and Arabic at the moment).

We are now actively looking for partners to expand the MBROLA and EULER projects.

Acknowledgments

We would like to thank early day contributors to this project: Piet Mertens for his MORLEX database and VERTEX part-of-speech tagger; Richard Beaufort for building a French phonetic dictionary (FRENCH-TCTS); people at Babel Technologies SA (Belgium) for their commitment to making tools freely available for non-commercial use, and Alan Black for his psychological support during the design of MLC, people at University of Aix-en-Provence for their encouragement at the beginning of the project.

6. References

Black, A.W., P. Taylor, and R. Caley, 1997. The festival speech synthesis system : System documentation. University of Edinburgh.

Clarkson, P.R. and R. Rosenfeld, 1997. Statistical language modeling using the cmu-cambridge toolkit. In *ESCA Eurospeech*.

Deroo, O., F. Malfrère, and T. Dutoit, 1998. Comparison of two different alignment systems: speech synthesis vs. hybrid hmm/ann. In *Proceedings of European Conference on Signal Processing*. Rhodes.

Dutoit, T., 1996. *An introduction to Text-to-Speech synthesis*. Boston: Kluwer Academic Publishers.

Dutoit, T., V. Pagel, N. Pierret, F. Bataille, and O. Van der Vrecken, 1996. The mbrola project: Towards a set of high-quality speech synthesizers free of use for non-commercial purpose. In *Proc. ICSLP '96*. Philadelphia, PA.

Hirschberg, J. and P. Prieto, 1994. Training intonational phrasing rules automatically for English and Spanish text-to-speech. In *Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis*.

Malfrère, F. and T. Dutoit, 1997. High-quality speech synthesis for phonetic speech segmentation. In *Proceedings of European Conference on Speech Communication and Technology*.

Malfrère, F., T. Dutoit, and P. Mertens, 1998. Automatic prosody generation using supra-segmental unit selection. In *Proc. 3rd ESCA/COCOSDA Workshop on Speech Synthesis*. Jenolan Caves, Australia.

Pagel, V., K. Lenzo, and A. Black, 1998. Letter to sound rules for accented lexicon compression. In *Proc. ICSLP '98*. Sydney, Australia.

Traber, C., 1995. *SVOX : the Implementation of a Text-to-Speech System for German*. Ph.D. thesis, ETH Zurich.

van Leeuwen, H.C. and E. te Lindert, 1993. Speech maker: a flexible and general framework for text-to-speech synthesis and its application to Dutch. *Computer Speech and Language*:149–167.

van Santen, J.P.H., R.W. Sproat, J.P. Olive, and J. Hirschberg, 1997. *Progress in Speech Synthesis*. Springer Verlag Edition.

Véronis, J., D. Hirst, R. Espesser, and N. Ide, 1994. NL and speech in the multext project. In *AAAI'94 Workshop on Integration of Natural Language and Speech*.