

FipsVox : a French TTS based on a syntactic parser

Jean-Philippe Goldman, Arnaud Gaudinat, Luka Nerima, Eric Wehrli

LATL - Department of Linguistics
University Of Geneva, Switzerland
{goldman,gaudinat,nerima,wehrli}@lettres.unige.ch

Abstract

FIPSVOX is a text-to-speech system for French developed at LATL. It is based on FIPS, a large-scale, multi-purpose, GB-based syntactic parser which produces detailed analyses and the MBROLA diphones-concatenation synthesizer. The syntactic information provided by the parser is directly exploited by the grapheme-to-phoneme module to handle heterophone homographs as well as French elision, denasalisation and liaison phenomena. The prosody generation module also uses this information to determine the dependency between phrases, the accentuation of syllables, and to identify particular syntactic structures such as extraposed constructions (cleft, heavy-NP shift, left-dislocation structures, etc.), and parentheticals to derive of appropriate prosodic patterns.

1. Introduction

The role assigned to syntax in prosody generation has varied a great deal over the last few decades. The main conclusion over these years seems to confirm the claim that the significance of a prosodic frontier increases with the significance of the syntactic one. But rhythmic constraints have to be considered, mainly for phonotactic reasons. Those two approaches can sometimes conflict, i.e. the prosodic outline given by the syntax doesn't agree with the rhythmic constraints. Fortunately, the second set of rules has a certain flexibility and an optimal trade-off can be found.

Current systems typically attempt to combine some shallow syntactic analysis with rhythmic principles. However, the naturalness of speech synthesised by such systems is still unsatisfactory due, in part, to the lack of a complete and accurate linguistic analysis.

Our system FipsVox [1] has a strong **linguistic component** (FIPS the syntactic parser) as a front-end. It is able to handle complex sentences and has some micro-grammars for special words like numbers (phone number, currency, time, date,...), acronyms, internet stuff (e-mail or URLs). Additional **phonetic** and **prosodic** modules take the advantage of rich linguistic information to compute the overall melodic curve, the phonetic sequence, and for each phone its duration. Finally, this whole information is sent to MBROLA for signal generation by diphone-concatenation.

2. Linguistic analysis

The linguistic analysis is done by the FIPS parsing system. It is a large-scale, multi-purpose, GB-based syntactic parser based on the Principles & Parameters Theory of Chomsky's Generative Grammar [2]. It produces detailed analyses which

are directly exploited by the grapheme-to-phoneme and the prosody generation modules.

The resulting syntactic structures contain lexical, phrasal, grammatical, and thematic information. The parser focuses on robustness, genericity, and deep linguistic analyses. Robustness is increased by the treatment of unknown words and the "no-failure" strategy of syntactic parses. Genericity is inherent to the Principles & Parameters Theory, where the Principles correspond to operations common to all languages. FIPS differs from shallow approaches in that it relies on detailed linguistic analyses, which clearly help to lexical, syntactic and semantic disambiguation.

3. Grapheme conversion

Word tokenisation and lexical analysis is a crucial step in the grapheme-to-phoneme conversion. Most speech synthesis systems have generally rules that are independent of the context or based on stochastic part-of-speech taggers. Our goal is to show how a global syntactic analysis can naturally disambiguate such problems of pronunciation.

The **grapheme-to-phoneme module** is based on the lexical database used by the syntactic analyser and an expert system for unknown words. This module also handles some phonological adjustments specific to French such as denasalisation schwa elision, resyllabation, liaison.

1. As a morpho-syntactic lexicon is a main resource of the syntactic parser, it is more natural to include the phonetic transcription within this lexicon (200'000 entries). Some semantic features and frequency information are also included.
2. Unknown words are phonetised by a classic rule-based system (about 700 rules), in which the rules are selected according to the graphemic sequence, graphemic left and right context, as well as syntactic context. This sub-module is also able to deal with :
 - a. any numbers (with particular rules ordinals, feminine numbers, phone numbers and numbers with units like time or currency)
 - b. acronyms (with a pre-processing that determines with phonological rules if it should be read or spelled)
 - c. proper names (with a pre-processing that guesses a possible foreign language and if so, apply assimilation phonetic rules)

Through the application of its grammatical rules and constraints the syntactic parser is able to solve most (but not all) lexical ambiguities likely to occur in a typical input text. In particular, heterophone homographs (e.g. words with identical spelling but different pronunciation) can be handled

successfully to the extent that they correspond to different word classes (a) or to a number distinction (b)

- a. 'président' can be a noun like in 'le président' (the president), pronounced [pʁɛzidɑ̃] or a verb like in 'ils président' (they preside) /, pronounced [pʁɛzidɑ̃]
- b. un fils (a son) [fils] / les fils (the sons or the thread) [fils]

A very small number of homographic ambiguities cannot be solved by syntactic methods and would require some level of semantic or pragmatic processing. This is the case, for instance, of the plural phrase 'les fils', which could be pronounced [lɛfils] 'the sons' or [lɛfils] 'the threads'.

In the next example (c), a local shallow syntactic analysis with a scope of few words (underlined) would be ambiguous about the word **content** as it can be read a verb in (d) or an adjective (e) :

- c. Les amis de l'enfant content et négligent la fin des histoires
- d. The parents of the child tell and fail the end of stories
- e. the child happy and negligent

According to its grammatical category, it is pronounced [kɑ̃] (as a verb) or [kɑ̃] (as an adjective). A mispronunciation would dramatically decrease intelligibility.

Moreover, for some others words like 'plus', 'tous', 'dix', 'six' or 'vingt', syntactic configuration is involved to determine the correct pronunciation. In (f), tous can not be read as a masculine determiner [tʊ] specifying the noun *journalists* (like in (g)) but as a floating quantifier [tʊ] related to the main clause (like in h) , because of a feminine gender agreement on the last word of the sentence *engagées*.

- f. Nous regardons tous les jeunes journalistes qui ont été récemment engagées
- g. (We look at all the young journalists that have been recently engaged...)
- h. (We all look at the journalists that have been recently engaged)

Notice that syntactic information is also useful for some of those phenomena. For instance, consider the case of liaison illustrated below. The sentences (i) and (j) are similar up to the sixth word but the liaison is absolutely forbidden in the first case where 'petit' is a noun functioning as a subject of the verb 'arbitre', while it is required in the second one, where 'petit' is a prenominal adjective modifying the noun 'arbitre':

- i. Le petit arbitre la rencontre qui a lieu ici
(the small one officiates the meeting that happens here)
- j. Le petit arbitre la rencontre à son appartement.
(the small referee meets her at her flat.)

The syntax can also solve some case of variable pronunciation of numbers according to the context like in (k),

where '31' could be pronounced in two possible ways (masculine: [tʁɑ̃] and feminine: [tʁɑ̃]). The juxtaposed noun 'journalistes' can not solve the ambiguity as it is both masculine and feminine. In order to verify the gender agreement, the system has to check the gender of the verb 'absentes':

- k. Les 31 journalistes sont absentes.
(the 31 journalists are away)

A large assessment of French phonetizer led to a good evaluation of our grapheme-to-phoneme conversion module.

4. Prosody generation

Here, we aim at computing a prosody that sounds as human as possible, i.e. naturalness, intelligibility, intra-speaker variation and a good reproduction of special prosodic patterns like focus, cleft, parenthesis,... The system has to determine the duration of each phone previously determined and an overall sequence of melodic targets.

Basically, our approach [3][4][5] relies on the syntactic dependencies of clitic groups (i.e. consecutive tools or function words followed by a meaning word) to determine prosodic or intonation groups. Then, several steps are necessary to determine the final melodic curve and the phoneme durations.

Creating clitic groups (or accentual group, henceforth CG) consists in finding the meaning words thanks to their grammatical category. This task may not be as easy as it appears. Prepositions, determiners are clearly function words, while nouns are not. But verbs can be split into different categories : auxiliary, modal, verb, past participle. Adjectives may not be accented if they are pronominal.

Once this is done, the concatenation all the previous function words and the selected meaning word gives a clitic group.

In our prosodic model, the intonational (or prosodic) group (henceforth IG) is a higher prosodic unit than CG and can group together one or more CG. The results should combine two generally accepted properties for IG grouping : 1. IG should have more or less the same length (in term of syllables) to satisfy basic phonotactic and eurythmic rules ; 2. The position and the strength of the prosodic boundaries that define IG, should rely on the syntactic dependency of constituents to avoid unfortunate associations of CG.

In order to determine IGs from the tree syntactic structure, a recursive algorithm starts at the syntactic root of the sentence and checks the number of syllables below this syntactic node. If it is greater than a threshold (typically 7 syllables per IG), each of the sub-constituents are processed separately, otherwise they are grouped into a single IG. This threshold mainly depends on speech rate.

Thus, several cases of attachment ambiguities can be solved by the parser thanks to syntactic features. A theoretical justification for these preferences is captured by Lonchamp's "right brother and uncle" (RBU) [6] rule, which states that a "right brother" constituent and its "uncle" constituent cannot constitute a prosodic group by themselves. They must either be combined with additional constituents or belong to distinct prosodic groups. The RBU rule has priority over rhythmic rules which favours well-balanced prosodic groupings in terms of number of syllables.

- l. Jean-François a acheté /des gâteaux à la crème

(Jean-François bought cream cakes.)

- m. *Jean-François a acheté/des gâteaux à la crèmerie
(Jean-François bought cakes at the dairy.)
- n. ?Jean-François /a acheté des gâteaux à la crèmerie
- o. ?Jean-François a acheté des gâteaux /à la crèmerie
- p. Jean-François /a acheté des gâteaux/ à la crèmerie

In (l), the rhythmic and the RBU rules are congruous (5/6 syllables). The rhythmic balance is correct in (m), but the RBU rule is violated, since the NP and PP are both sub-constituents of the VP node. As we noted above, according to the RBU, they can either be grouped together with the verb in a single prosodic constituent, as in (n), which violates the rhythmic rule (3/10 syllables), or belong to distinct prosodic groups, as in (o) or in (p). Since (o) does not have well-balanced constituents (9/4 syllables), (p) (3/6/4 syllables) appears to be the optimal solution with respect to both constraints.

Once the IG are determined, each syllable is assigned an accent type (unaccented, final low accent, final high accent, focus accent) and an optional following pause according to parameters like: the distance to the next prosodic boundary, the strength of this boundary, the position of the syllable in the IG and in the sentence, the length of IG. Some rules are set to avoid accent clash as well as long sequence of unstressed syllables.

The register is the macro-prosodic frame of the melodic curve at the sentence level. It is defined by two lines : the baseline and the high line. The declination effect is modelled by a decreasing and narrowing register. Each syllable accent is converted into a relative height within the register (i.e. unaccented syllables are set on the baseline, accented syllable are set above, below or on the high line, according to the accent type and some others parameters).

Additional adjustments are done at all levels of the prosodic model. For instance, parenthesis has an effect of lowering or raising the register. Some rules modelling a focus due to cleft or extraposed elements, affect each syllable of the focused constituent as well as the surrounding syllables.

As an example, let's see how cleft constructions are prosodically marked. Although they share superficial similarities with relatives clauses, they receive a distinct prosodic pattern, with high stress on the focus (the cleft element). Example (q) is pronounced differently as an answer to question (r) or to question (s). In the first case it is cleft, in the second a relative clause.

- q. *C'est le lapin que j'ai adopté.*
(It is the rabbit that I adopted)
(This is the rabbit I adopted)
- r. *As-tu adopté le hamster ou le lapin ?*
(Have you adopted the hamster or the rabbit?)
- s. *Qu'as-tu dans les bras ?*
(What do you have in your arms ?)

Finally, syllable duration is computed on the basis of the accent type, the prosodic boundary strength, the position within the IG and the number of phones of the syllable. A repartition model assign the duration of each phone of this syllable depending on the syllable duration and default mean

and standard deviation of durations of phones measured in a training corpus.

Thus, each phone is assigned a duration and a melodic target. These information are collected and written in a standard MBROLA format for the final speech synthesis [7].

5. Current developments and future work

The generic aspect of the linguistic theory of the FIPS parser and its object-oriented implementation allow an easy development towards multilingual application. In other words, more languages should be added soon to the text-to-speech system with a complete linguistic component.

Moreover, the numerous foreign passages in current texts like e-mails or news require a good linguistic analysis of multilingual sentences (i.e. part of texts with some language codeswitches). One can think of foreign proper names but longer code switches are usual:

- titles of books, songs or movies
- complete proper names
- quotes
- society or organisation names

A current project aims at defining the nature of such codeswitches and how they interact linguistically with the main sentence. As an example, the pronunciation of the French citation of the famous movie '*La cité des Anges*' in an English text crucially needs a liaison between the preposition '*des*' and the noun '*Anges*', that is to say phonetised as [lɑ̃sɪtɛʁɑ̃ʒɑ̃ʒ] instead of [lɑ̃sɪtɛʁɑ̃ʒɑ̃ʒ] which would produce a hiatus and a lack of intelligibility.

An undergoing project focuses on improving intonation with corpus-based melodic rules. Furthermore, multilingual parsing is currently developed.

6. References

- [1] Gaudinat A., Goldman J-P., Wehrli E. '*Le système de synthèse FIPSVox:syntaxe, phonétisation et prosodie*', XXIIèmes JEP, 1998, Switzerland, pp.139-142.
- [2] Chomsky, N. & Lasnik, H.: "The Theory of Principles and Parameters" in Chomsky, N. The Minimalist Program, Cambridge, MIT Press, 1995, 13-127.
- [3] Goldman J-P. & E. Wehrli, '*Deriving prosodic patterns from syntactic structures : the case of extraposition, clefts and extraposition*', ESCA workshop on Intonation : Theory, Models and Applications, eds Botinis A., 1997, Athens, Greece, pp.153-156.
- [4] Mertens P. '*Un algorithme pour la génération de l'intonation dans la parole de synthèse*', Actes Conférence TALN 1999, pp.233-242.
- [5] Mertens P., Goldman J.-P. et al. '*La synthèse de l'intonation à partir de structures syntaxiques riches*' TAL 2001, Toulouse, France
- [6] Longchamp, F. '*Prédire l'intonation d'une phrase affirmative*', journées ATALA, Paris, fév. 1996.
- [7] Dutoit T. '*The MBROLA Project*' Info. available on <http://tcts.fpms.ac.be/synthesis/mbrola.html>