LINGUISTIC-PROSODIC PROCESSING FOR TEXT-TO-SPEECH SYNTHESIS IN ITALIAN

Philippe Boula de Mareüil

Ean Informatique a Lernout & Hauspie Company
4 rue Jean Rodier — F-31400 Toulouse
Tel.: +33/0 5 61 36 07 77 — Fax: +33/0 5 61 36 89 11
e-mail: mareuil@elan.fr — URL: http://www.elan.fr

1. INTRODUCTION

The current Elan text-to-speech (TTS) synthesis system comprises eight languages: French, Spanish, German, British English, American English, Russian, Brazilian Portuguese and Italian. In this paper, the linguistic processing applied to the phonetic-prosodic transcription of (standard) Italian is described, for the purpose of TTS synthesis. It proceeds in 5 steps: tokenisation and normalisation of abbreviations, numbers, etc.; part-of-speech tagging, based on function words, terminations and contextual heuristics; shallow parsing, based on a chunk grammar; grapheme-to-phoneme conversion, lexical stress assignment and syllabification by an expert system, particular attention being paid to loanwords and proper names; at least, prosody generation. The last three steps are emphasised.

The general organisation of this article roughly traces the order of the program’s operations, before the concatenation of units (namely, diphones). Though, we will not focus on the pre-processing linked with typographic conventions — work is in progress, on the pronunciation of acronyms in Italian. Nor will we detail the disambiguation of words in -anol-onol-ero and in -intro, which may be conjugated verbs or non-verbs: negative heuristics, of a distributational character, are applied, which stipulate, for instance, that one cannot have a conjugated verb after a determiner. Indeed, this is important for lexical stress assignment, together with the bi-partition between content words and function words — clitics, which are unstressed, unless they terminate a sentence or precede a punctuation mark.

More attention will be paid to the last three steps: phonemic transcription, syntactic analysis and prosody (accentuation and intonation). The overall system is under evaluation on a 200,000 word corpus. On a text and a lexicon (of about 1,500 words each), the accuracy rate of pronunciation on words happens to range between 95 % and 99 %.

2. PHONETIC TRANSCRIPTION

Grapheme-to-phoneme conversion may be regarded as relatively straightforward and requiring few rules, compared to languages such as English or French. Cases of one-to-one correspondence between a grapheme or a graphemic string and its phonemic transcription are the majority. Nonetheless, the linguistic history of the Italian language, in comparison with these other languages, shows that it is characterised by a strong and vital presence of regional and local dialects, as well as by a noticeable difference between these varieties. Herein, the pronunciation norm we adopted is Toscan (fiorentino colto).

2.1. Phoneme set

The first question is that of the phoneme inventory and of their code. In Europe from now on, the phonetic alphabet developed within the framework of the SAM projects [18], SAMPA, is a well-established standard for the natural language and speech processing community. Let us say here that he notation adopted for geminates consists of doubling the first symbol of the consonant: hence /hS/, /dZ/, /tS/, /dz/ for the affricates.

We dealt with these geminates and all the others as two segments. Although the phoneme /S/ exists in Italian — unlike the lone /Z/ — it does not seem natural for a native speaker to record the affricate /hS/ as /h+s/. We therefore incorporated the three phonemes /S/, /S/ and /dZ/ in the database.

We kept the glides /j/ and /w/, which are shorter than the corresponding vowels /i/ and /u/ [24]. Their phonological status is questionable, but the existence of minimal pairs such as /kwil/ (qui) ~ /kwi/ (cui) and /allojamo/ (from alleviare) ~ /allojamol/ (from alleviare) is enough to have them considered as phonemes and not as mere combinatorial variants. In return, we did not duplicate the vowels in a stressed version and an unstressed version, even though the stressed [a] has a more open timbre. The 30 phonemes we come up to is the figure on which most works agree: /p, t, k, b, d, g, m, n, j, l, S, s, tS, dZ, ts, dz, l, r, s, z, f, v, a, e, i, o, O, u, j, w/ in SAMPA.

2.2. Grapheme-to-phoneme conversion

Classical issues of grapheme-to-phoneme transcription in Italian, such as ‘c’ and ‘o’, ‘s’ and ‘z’, ‘i’ and ‘u’, ‘c’ and ‘g’, were addressed, following a knowledge-based approach. Particular attention was also paid to loanwords and proper
names, where a conflict appears between respecting the original spelling, and approximating the original pronunciation, by following the Italian conventions.

2.3. Lexical stress assignment

A crucial problem in Italian is the determination of the lexical stress position. It is often assumed to be “unpredictable”. Let us mention, though, that only verbs may be stressed on the fourth or fifth syllable from the end, and that 80% of the words are stressed on the penultimate syllable [29]. For the remaining ones, the observation that words sharing the same termination (e.g. -glibse, -gico, -gine, -giglio, -giore, -gione, -gione, -gione, -gione) often show an identical accentual scheme suggests a set of morphological rules. The concept of “suffixoids” (i.e. pseudo-suffixes) has been put forward [11]:

- **strong** suffixoids (i.e. always attracting the stress);
- **weak** suffixoids (i.e. always shifting the stress onto the previous syllable);

A second set of rules is devoted to irregular verbs of the 2nd conjugation, in -glibere, -dere, -ere, -rompere, -crovere, etc., which are stressed on the antepenultimate syllable in the infinitive [12]. Moreover, a few regular verbs of the 1st conjugation shift their stress one syllable backward in the singular and in the 3rd person of the plural in the indicative and subjunctive present. E.g.: *anima* → /anim9/, *animano* → /animano/; we took into account. These verbs generally stem from a noun or an adjective stressed the same way.

When there is pronon enclisis with the infinitive, the stress location rises no difficulty with simple pronouns. But we may also have forms like *parlare*. This was processed, as well as the accentual pattern of forms such as *cercarmelo* and gerunds — which obey more general rules. The cases of imperative and absolute past participle, with simple pronouns, generate more ambiguities, particularly with the pronouns *si, le, li, ne*, which we did not try to handle on the whole.

2.4. Methodology

We split the lexical stress assignment into two procedures which are linked: the first one is devoted to stresses which may be obtained directly from the spelling, and which should be kept in memory during the grapheme-to-phoneme conversion. The second one uses this result. Two *flex* files are associated to them: each one contains about 600 rules composed of two main parts, which are partially ordered — this number is only indicative, since the rules may be very compact (right contexts are especially widely used).

Communicating the same way, a third file of rules was added for syllabification: it comprises about 15 rules which set syllable boundaries between a vowel and a consonant followed by a vowel, between two /s/, etc. Unlike in French or in English, where a consonant cluster like /st/ would be split up, the syllabic break falls in Italian before the “impure s” — hence the syllabation *que-sto*. Very controversial (see e.g. [25]), this “tautosyllabic s” with respect to the next syllable” is at least what is advocated by traditional grammar.

So, we have three finite-state machines in cascade, each layer being defined by regular expressions. The output is then connected with the next module.

3. APPLICATION OF A CHUNK GRAMMAR

Chunk grammar [1], which merely consists of splitting sentences into non-recursive segments, partially draws its inspiration from studies in psychology about pause duration, in reading, and about “naive” parsing of utterances. Evidence for the existence of these chunks is the so-called *raddoppiamento* (fino)stintattico, in central and southern Italian [14]. The present analysis is not exhaustive, but results in a quick, deterministic and robust parsing — on running texts —, like the *chinks ‘n chunks* algorithm [21] which is used in the Bell Labs TTS system. Likewise, in the CSELT system ELOQUENS®, the position and the type of prosodic boundaries are determined from the grammatical tags of words and from rhythmic constraints [19].

Bolinger [7] namely speaks of information *chunks*, defined by intonational groups. *The typical chunk consists of a single content word surrounded by a constellation of function word* [1]. We distinguished four types of chunks:

1. a sequence (possibly empty) of function words and one or several content word(s) preceding a function word;
2. a sequence (possibly empty) of function words and a word preceding a weak punctuation mark;
3. a sequence (possibly empty) of function words and a word preceding a final period, a semicolon or an exclamation mark;
4. a sequence (possibly empty) of function words and a word preceding a question mark.

Like in constraint grammars (e.g. [20]), our analysis is based on local constraints; and linguistic descriptions are encoded by labels and not a phrase-structure. A new type of label is added to the part-of-speech attached to each word: by default 0, the last word of the chunk indicates the type of chunk. E.g.:

Guglielmo non beve perché è infelice.
1 0 1 0 0 3
Deve essere all’aeroporto per le sei e trenta.
0 1 0 1 0 0 0 0 3
Quando sono arrivati i messaggi di Francesco?
0 0 1 0 1 0 1 0 4
On the Italian Multext corpus [9], the most frequent chunk in terms of parts-of-speech, PREPOSITION NOUN, is summarised as 01. And the 10 most frequent types of chunks cover more than 86% of the occurrences.

Even though the link between syntax and prosody is controversial (e.g. [10, 7, 21, 2, 22]), such information is used, together with the stress location, the type of stressed syllable (open/closed, followed or not by a geminate) and the type of phoneme (voiced or unvoiced consonant, stressed or unstressed vowel), for the generation of prosodic parameters and the distribution of pauses. An inductive (bottom-up) approach was adopted, this way following works by the “Dutch school” (IPO). Building the prosodic model can be seen as a clustering task, aiming at making the problems discrete by categorising them. This was done on the basis of a male voice selected from the Multext corpus — the speaker read 15 passages of about 20 seconds (5 sentences linked by a coherent thematic structure).

4. PROSODY GENERATION

4.1. Pitch

The pitch of each phoneme is defined by two values: initial and final heights. For a male voice and a female voice, a baseline and a topline are defined, as well as different pitch levels (¼- final heights. For a male voice and a female voice, a baseline

The pitch of each phoneme is defined by two values: initial and final pitches. For a male voice and a female voice, a baseline

4.2. Duration

For duration, multiplicative coefficients are applied to the intrinsic values of each phoneme (see table 1). Six factors are distinguished, in particular examining the possible presence of a following pause, to account for the pre-boundary lengthening. The latter, which has been shown for a number of languages, could indeed be universal [9, 26]. Considering Italian as a syllable-timed language — unlike e.g. English, which tends to exhibit isochronous iter-stress intervals —, the reduction of unstressed vowels was avoided [16]. Nor was any lengthening applied on the consonants of stressed or final syllables: too short, it would not be perceived — the just-noticeable difference being 20 ms, according to [5] —; too long, it would risk to have a geminate heard, instead of a simple consonant. Even the lengthening of the reinforced palatals, /L/, /S/ and /J/, (almost) geminated in intervocalic position, is implicitly captured. Therefore, this phoneme-based model seems to be more suited than others, based on units of higher level [8, 4] — or at least sufficient, for a couple of reasons:

Arguments exist in favour of both approaches [17], and this old antagonism is actually a false problem: the most interesting issue is that of the alignment with the units which constitute the associated segmental string [13]. Now in TTS synthesis, applying a prosodic structure to a text is precisely what must be done, and what is done here.

Table 1: examples of mean durations (plosives).

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Nevertheless, we agree that possible improvements could benefit from considering:

- the size of the word;
- the stress position with respect to the beginning of the word [28];
- the type of word (e.g. function word);
- the type of the consonant following the vowel (e.g. /z/ vs voiceless plosive).
4.3. Pauses

Pauses are associated to punctuation marks: of one second for strong punctuation marks (which are used as formal limits for silences which are greater than a certain threshold. So, work is in progress to see if we can extract regularities: if pauses happen to appear between such and such grammatical categories, after certain words...

5. CONCLUSION

Various features traditionally assumed about the Italian (spoken) language have been addressed and automated. A modular approach was adopted, integrated in a common framework called SYC [6], which guarantees portability under different versions of compilers and of operating systems — whatever the byte order. The program was written in flex and in ANSI C (with associative tables).

We hope that the results provided by the (diphone-based) TTS synthesis, with a simplified representation of syntax and prosody, will help the study of the correlations between these two poles. Unlike other systems such as CSELT’s [3], which uses stochastic methods derived from speech recognition, we privileged a structural approach, rule-based and non-lexicalist. The originality of this work lies in a unified use of finite-state transducers. An on-line demo is available through the Internet (http://www.elan.fr).

Together with language identification (based on the 200 most frequent words of French, English, German, Spanish, Portuguese and Italian), the Elan e-mail reader service, Dial & play, has also been extended to Italian.

6. REFERENCES