On the Pronunciation of Acronyms in French and in Italian

Philippe Boula de Mareüil & Franck Floricic

1 Ean TTS
4 rue Jean Rodier, F-31400 Toulouse
mareuil@elan.fr

2 ERSS, Université Toulouse-Le Mirail
5 allées Antonio Machado, F-31058 Toulouse
floricic@univ-tlse2.fr

Abstract

Acronyms are used more and more nowadays: this article describes their pronunciation in French and in Italian. Rules are proposed, so that a text-to-speech converter may know whether it must spell or read acronyms, and which way. Our analysis, which gave rise to a computational realisation, leans on the number of letters which constitute acronyms, on the vowel/consonant opposition, as well as on the distribution between continuous and momentary consonants.

1. Introduction

Whether in French or in Italian, acronyms are particular cases of abbreviations which are used for shortening groups of words, by keeping their lone initial letters or syllables, possibly regardless of determiners, prepositions and/or conjunctions (e.g. LCI for La Chaîne d’Information, in French, IVA for Imposta sul Valore Aggiunto in Italian). These acronyms are used more and more in contemporary running texts, to designate administrations, organisms (even the smallest ones), companies, scientific concepts, sport clubs, countries, instruments, etc. (e.g. VTI, UNESCO, IRC, FIOM). They even get a legal value, and may be protected as trade marks [1]. Whatever the newspaper page you happen to read, you are immediately confronted to acronyms, and writers no longer disdain to have resort to them — particularly in French. In Latin, expressions such as SPQR (Senatus Populusque Romanus) were already customary, and the French language exhibits similar forms as soon as the XVIth century (with RPR for Religion Prétendant Réformée). Sometimes, nobody remembers that a word is or used to be an acronym: this is the case for LASER (or laser, for Light Amplification by Stimulated Emission of Radiation), RADAR (or radar, which also looses its initial upper-case letter, unlike the following ones, for Radio Detecting and Ranging), RAI (or Rai, for Radio e Audizioni Italiane), FIAT (or Fiat, for Fabbrica Italiana Automobili Torino), ALFA (or Alfa, for Anonima Lombarda Fabbrica Automobili, which gave alfetta by suffixation).

Facing an acronym, two attitudes are possible: to read it as if it were a word (e.g. CEE, CIRASS), or to spell it (e.g. ICQ, ITT, MBA, SOS). Italian people, unlike English people, prefer the first tendency [2]. In French, it is more complicated [3, 4] — almost fifty-fifty. The acronym ONU, which receives a couple of possible orthisations, testifies of it.

To us, the interest of studying how to pronounce acronyms is twofold: on the theoretical ground, first, we handle the basis of the phonological system — even though that of acronyms is incomplete (n/u, for example, seems to be missing in both French and Italian); and we directly test the importance of phoneme clustering. On the practical ground, then, it is difficult to recognise PDU, when a text-to-speech (TTS) synthesis reads /pdy/ or /pdu/. Still, few rules have been applied concerning this subject: in the VoxBox TTS system [5], for instance, only rules apply to acronyms which do not comprise vowels (e.g. RG), which are logically spelled out. Next sections are devoted to current theories and corpora (Section 2), and to the analyses of French (Section 3) and of Italian (Section 4).

2. Current theories and corpora

Some people speak of acronyms when a syllabic pronunciation is possible, and of sigle for gatherings of letters which do not constitute syllabic words and result in an alphabetical pronunciation. As there is no consensus on these definitions, we will not make this distinction. Typography consisting of having a dot follow each letter in spelled acronyms — which is no longer necessary in real acronyms — could also solve the ambiguity. But it is not reliable.

A variety of factors may account for the pronunciation of acronyms: the familiarity one has with them, the contempt one wants to show for what they represent, or the entertaining effect sonorities can produce. Current attempts to explain these mechanisms are very interesting on the linguistic field, but our purpose here is to find rules which may indicate if a TTS system must read or spell a given acronym, with an exclusive or, because we do not introduce variability: when an acronym has two realisations, we have not searched which one is the more frequent, but kept the one which obeys a more general rule. Now, the following rules are not sufficient,
Fore being read, acronyms need comprise a group CV (C=consonant, V=vowel), which indeed constitutes the most natural syllable [6]. However, the very acronyms in CV are more often than not spelled out, which multiplies hiatuses, especially if C is a plosive (e.g. BO, CA).

Stating that the reading of an acronym should lead to a form composed of at least three phonemes is not a more sufficient condition, which is also the problem with approaches based on moraic phonology, whether for French [7, 8] or for Italian [9]. Pieniat proposes minimum and maximum weight thresholds for an acronym to be read in French, and provides examples of possible conflicts [8]. The minimum threshold of two more (corresponding to a monosyllabic word with branching rhyme or to a disyllabic word) defines a limit below which an acronym is compulsorily spelled; and the maximum threshold of three syllables defines another limit above which acronyms are preferably read. These syllabic weight constraints coexist with a set of structural constraints which determine the pronunciation of acronyms whose constituents may be considered as ill-formed. Yet, they are not sufficient: OPA is spelled in French (not in Italian).

The acronym gender (which is apart from a few exceptions that of the head of the noun phrase), the domain (e.g. teaching), the spelling of the letter (we say /v/ but /ɛ/ in French, /w/ but /ɛɛ/ in Italian) do not seem to play a role either.

In order to let the speech synthesis know how to utter acronyms, we wrote a few rules on the basis of a couple of corpora:

- of about 1,000 acronyms in French [8] (see Figure 1);
- of more than 800 acronyms in Italian [9] (see Figure 2);
- with the number of letters of each acronym, its meaning and its pronunciation(s).

![Figure 1. Repartition of French acronyms in relation to the number of their letters.](image1)

![Figure 2. Repartition of Italian acronyms in relation to the number of their letters. It can be observed that the most numerous acronyms are composed of 4 letters, against 3 in French. Acronyms being of a less frequent use in Italian, this illustrates the fact that we have resort to them when the words they abbreviate are really too long.](image2)

Our analysis relies on the number of letters which acronyms are composed of, as well as on phonotactic constraints. The first criterion is purely orthographical, but what is an acronym, if not a play on alphabetical writing? Nor is it as far from phonology as it may look at first sight, since acronyms are first of all a written abbreviation, a visual language, an economy of typography [10].

Therefore, our program begins with evaluating the length of acronyms. Given these ones range from 2 to 9 letters (seldom reached maximum), they can be visualised as a whole, and we may think that man does the same.

### 3. Analysis of French

In French, the principle of linguistic economy — limiting the spent energy while maintaining intercomprehension — only applies beyond 3 letters (that is the most frequent acronyms); indeed, since we find it harder to remember them, a memory problem is raised. Hence the following rules, where the term "continuous consonants" corresponds to the set of letters F, H, J, L, M, R, S, V, W, X, Z:

Acronyms are spelled out by default, but:

- acronyms in CVC are read if V is a vowel (e.g. CAC, ZEP, where the homonymy with ZUP may also intervene);
- acronyms in VCV are read if C is a continuous consonant other than H (e.g. IRA but API, HIV).
• acronyms in VCVC, CVVC, CVCV, VVCV, VCVV (likewise, C(C)VC, CVC(C)) are read if C₁, C₂ or C₃ is a continuous consonant)¹ (e.g., SIDA, SMIC, which gave sidēen, smicard, respectively);

• acronyms in V₁CVC₂ are read if V₂ is a vowel other than E (e.g., ONVI, which gave onniologie, but OCDE — the elision of the final E would lead back to a pattern VCC, which should yield to spelling —);

• acronyms of at least 5 letters are read if they comprise a vowel (e.g., SOCRATE) — otherwise, it would be too long to spell them.

Once the partition between read acronyms and spelled acronyms is made, the problems are not completely solved. As far as the spelled pronunciation is concerned, indeed, there is little choice, even though:

• the Toulouse Football Club is called /tufen/;

• E may be oralised /æ/ or /e/ (e.g., EOR) and W as /w/ (e.g., WC);

• GI, DJ, CB are sometimes pronounced in the English way (hence cibiste);

• the tendency to say 2E for EE, 3A for AAA, etc. seems to gain ground, even if it is only encountered with CH, in the base corpus.

But the reading of acronyms differs from the normal reading in several points:

• final consonants are pronounced, which is not always the case with speech syntheses for -d, -g, -s, -i, -y (even doubled), -t, -s (e.g., DBAAS, CROUS);

• -ER is pronounced /er/ and not /e/ (e.g., MAIZE);

• groups of the form vowel + N or M are not nasalised (e.g., CNAM, SAGEM, LIMSI);

• the letter E is pronounced /e/ in open syllable within read acronyms, even though it stands for European (e.g., ECU) or emploi in ASSEDIC. So, our program examines if the E is followed by at least two letters, and if the second one is an L, an R or a vowel rather than a final E. In comparison with a whole syllabification, this is enough for acronyms. This way:

1 Note we counted C and G among momentary consonants (UCI and DGER are spelled out, although their spellings (œel and /œl/ start with a continuous phoneme, and although they themselves may be pronounced /œ/ and /œ/ within read acronyms respectively, before E or I: e.g., DAGES, CILF /silf/ (which stands for the Conseil (fr.) International de la Langue Française.

Particular attention should be paid to acronyms in EN- — often meaning Ecole Nationale (e.g., ENAC) — so that the synthesis does not read /en/ as in enivrer. Facing this issue of desenuisalisation, Yvon, on the basis of a 3,000 acronym database, dissociates the behaviour of E and I on the one hand, of A and O on the other hand (more often nasalised), and explains the pronunciation /gæ/ of GAN by the fact that /gæ/ would be too light [11]. In a framework which draws its inspiration from Optimality Theory, he proposes an additional constraint to choose amongst all the possibilities, aiming at having the oralised form as close as possible to the written form, so that the latter may be rebuilt without ambiguity [12]. But we dispose of too few acronyms to confirm this tendency, and in the case of OTAN, there is even hesitation. Anyway, the implementation of the rules above yields to an estimated accuracy rate of 96 %, on Plélat’s corpus. Most of the errors stem from acronyms which should be spelled, and which are not, according to our rules (e.g. UEFA, SVP, VIP). Quite often, their pronunciations may be explained by analogy phenomena (e.g. DUT with IUT), or by the presence of unusual consonant sequences (e.g. FNSEA); hence the approach to the Italian language.

4. Analysis of Italian

In Italian like in French, the principles which impose the choice between reading and spelling interact in such a way to yield what could be identified as the most probable output candidate. Therefore, it is worth pointing out that a unique rule cannot account for the complex phonological behaviour of Italian acronyms. Acronyms of two letters are regularly spelled out. On the other hand, when an acronym is composed of more than two letters, it is possible to identify a set of conditions on spelling which function as filters. In other words, the following parameters have the property of blocking reading and imposing spelling:

• the presence of one of the characters H, J, K, W, X, Y, which are not typically Italian (e.g. TWA, UKY);

• the presence of a pair of consonants such as -CV, -BM, -BC, -FC, -CG, -DN, -FB, -MD, -VG, -DV, -PD, -FC, -GC, -PM, -CF, -SR, -DT, -DS, -IP, -GB, -GM, -DP;
• the presence in onset of a pair of consonants such as -LS-, -MS-, -CM-, -CN-, -CT-, -FF-, -RN- (e.g. CNCA)²;

• the absence of vowels (e.g., CNR).

Other exceptions are foreseen. For the lexical stress assignment, some rules specific to acronyms have been added, for instance stressing the initial syllable of acronyms in (C)VCVCVC (e.g. UNICEF). Note that 84% of acronyms are stressed on the first syllable. These rules suffice to properly phonemise 83% of the corpus of acronyms. If only binary spelling/reading decisions are kept (without counting mixed oration modes such as PltUP /p'ldwp/), 97-98% of the decisions are correct.

5. Conclusions

In this paper, our goal was to improve the quality of grapheme-to-phoneme conversion, which is indispensable for TTS synthesis, and which is also useful for spelling checking, information retrieval and automatic speech recognition systems. We focussed on the pronunciation in French and in Italian, of “words” which are not (or little) represented in dictionaries, and which have remained marginal in phonology so far, whereas they proliferate in real-world texts — indeed, few constraints are involved in their construction.

These problems, which are isolated, heterogeneous but important, must be solved for TTS synthesis [13]. It is also interesting to study if the results obtained apply to the pronunciation of e-mail addresses and URLs, as well as to other languages like English, German and Portuguese, as studied by Trancoso [14]. Work is currently in progress in this field, on English.

In our program, the letters of acronyms to be spelled are separated, taking care of apostrophes and other punctuation marks such as (, :, " ), which may precede or follow. An acronym is detected as a word entirely written with capital letters — or combining capital letters (possibly separated by dots) and figures.

The major problem, which we did not solve completely, obviously comes from other words written with capital letters (rules especially). We then decided to avoid the spelling out of function words such as LE, DE, in French or DI, IL, HA, CHE, in Italian. This is a trade-off, but finer detection of acronyms requiring part-of-speech information, would be more helpful.

² It is as if the decision were governed by the onset: it cannot be too complex, and the second part of the onset must be more sonorant than the first part, following the so-called Sonority Sequencing Principle. Inversely, we may have a complex coda (e.g., RST). The end of acronyms is also truncated more easily than the beginning.