Evaluation of grapheme-to-phoneme conversion for Text-To-Speech synthesis in French

P. Boula de Mareüil (LIMSI), F. Yvon (ENST), C. d'Alessandro (LIMSI), V. Aubergé (ICP),
M. Bagein (TCTS-FPMS), G. Bailly (ICP), F. Béchet (LIA), S. Foukia (LPL), J.-P. Goldman (LATL),
E. Keller (LAIP), D. O'Shaughnessy (INRS), V. Pagel (TCTS-FPMS), F. Sannier (ICP), J. Véronis (LPL),
B. Zellner (LAIP)

LIMSI-CNRS, BP 133, 91403 Orsay cedex, France
ENST, 46 rue Barrault, 75013 Paris, France
ICP, CNRS ESA 5009, DU, 1180 avenue Centrale, BP 25, 38040 Grenoble cedex 9, France
INRS, 16 Place du Commerce, Ile-des-Soeurs, Verdun, Québec, H3E 1H6, Canada
LAIP, Faculté des Lettres, Université de Lausanne, 1015 Lausane, Switzerland
LATL, Université de Genève, 2 rue Candolle, 1211 Genève 4, Switzerland
LIA, Université d'Avignon, 339 chemin des Meinajaries, BP 1228, 84911 Avignon cedex 9, France
LPL, CNRS ESA 6057, 29 avenue Robert Schuman, 13621 Aix-en-Provence Cedex 1, France
TCTS, Faculté Polytechnique de Mons, 31 boulevard Dolez, 7000 Mons, Belgium
listeb3@limsi.fr

Abstract

This article reports on a cooperative evaluation of grapheme-to-phoneme (GP) conversion, for Text-To-Speech (TTS) synthesis in French. This work has been carried out in the framework of a general evaluation of various speech and language processing devices, conducted under the ægis of the French-speaking Francil network. The methodology and the corpus are described. The results of 8 systems from France, Belgium, Switzerland and Canada are analysed. They give a fairly accurate picture of the progress achieved, of the state-of-the-art and of the problems still to be solved, in the domain of GP conversion for French. Moreover, the resources and data will be put at the disposal of the scientific and industrial community, in order to be re-used in future benchmarks. On the other hand, the methodological issues have been extensively discussed, and possible improvements have been envisaged.

1. Introduction

This paper reports on a cooperative evaluation of grapheme-to-phoneme (GP) conversion, for Text-To-Speech (TTS) synthesis in French. This work has been carried out in the framework of a general evaluation of various speech and language processing devices, conducted under the ægis of the French-speaking Francil network. The methodology and the corpus design are described. The results o 8 systems from France, Belgium, Switzerland and Canada [1] are analysed (Aubergé, 1991; Béchet & El-Bèze, 1996; Boula de Mareüil, 1995, 1997; Dutoit, 1993; Gaudinat & Wherli, 1997; Keller, 1997; O'Shaughnessy, 1984; Yvon, 1996). The systems involved all rely on a rule-based approach: the rules, whose number ranges between 500 and about 4,000, may be superseded by look-up in an exceptions lexicon (up to thousands of word forms). However, they differ in the amount of linguistic processing, as far as syntactic analysis and specific modules for pronouncing proper names are concerned.

TTS systems usually involve three main stages: GP conversion, prosody assignment and signal production. The experiments we present here only concern GP conversion, which is a difficult issue in French (like in English): besides morpho-phonological ambiguities, heterophonous homographs, the mute *e* (or schwa, which may be uttered or dropped), glides (which may be pronounced a syllabic way or not), laisons (realisation of final consonants, in the context of a following word initial vowel), complex problems are raised by *extra-lexical* items such as proper names, numbers and abbreviations. which are frequent in real-world texts.

Quite an abundant litterature exists about TTS evaluation methods (Silverman *et al.*, 1990; Carlson *et al.*, 1990; van Bezooijen & Pols, 1990; van Santen, 1993; Kraft & Portele, 1995; Sorin & Emerard, 1996; Klaus *et al.*, 1997; Benoît, 1997; Pols & Jekosch, 1997). To the best of our knowledge, though, our work represents the very first attempt towards the definition and use of an objective evaluation methodology for GP conversion in French. In this project, the evaluation methodology is tuned to the framework of TTS systems, which usually output a single phonemic string for each input. More precisely, we dealt with book or newspaper reading: running texts rather than lexica.

In the next section, we present the methodology. Section 3 describes the corpus. Section 4 presents the results obtained. Section 5 discusses the methodology. Section 6 summarises the results; directions for future studies are also envisaged.

1. LIMSI (Orsay), ENST (Paris), ICP (Grenoble), LIA (Avignon), INRS (Québec), TCTS (Mons), LAIP (Lausanne), LATL (Geneva). LPL (Aix-en-Provence) was in charge of organisation and corpus development.

2. Methodology

At first glance, evaluating a GP conversion device looks quite straightforward. But a reference problem is raised, because speech is subject to variability along multiple dimensions (stylistic, idiolectal, etc.). Even though most dictionaries consider that there exists one and only one acceptable pronunciation for each word of the language, the ideal situation where every one would do exactly the same phonetic distinctions is far from reality (Martinet, 1960). Additionally, the GP transcription module often interact with other modules of the TTS system: a possible consequence is that a divergence from the reference does not impair the comprehension of the output speech.

Despite these limits, a common phonetic alphabet was defined: it is very close to the well-known SAMPA (Gibbon, 1997), which has already been used in the context of TTS assessment. The participants were then asked to stick to it. As for encoding variants within the reference corpus, the solution that was adopted amounted to limitting this one to a lattice of pronunciations. An example, coded in SAMPA (the | symbol separates possible alternatives), is:

Entre le monde et lui, la guerre est ouverte .
A~tR@ l[@|] mO~d e lHi lA gER [e|E][t|] uvERt[@|].

Such a format can easily be handled at the computational level, and does not make any assumption regarding the scoring measures that is applied during the evaluation phase. Furthermore, this representation of phonological variants can easily be extented, to accommodate for additional variants in an incremental manner.

Stemming from the experience of previous works, the evaluation protocol consisted in submerging, in a much larger corpus (12,000 sentences in our case), the text on which the results were analysed. This portion was of course secretly selected by the organiser (LPL, Aix-en-Provence). The task given to each participant consisted of phonetising the entire text within a restricted time frame. A dynamic programming algorithm was used, to align the phonemic outputs with the reference.

3. Corpus

After a preliminary test phase which enabled us to verify the validity of our approach (d'Alessandro *et al.*, 1997), a second phase took place during the summer 1997. The organiser of the test campaign hand-transcribed the test corpus, which consists of articles extracted from the French newspaper *Le Monde* of January 1987; the decision was made to work on this kind of text, because it contains a full assortment of the typical difficulties for GP conversion. About 2,000 sentences were selected, with the specific concern that numbers, (foreign) proper names and acronyms should be significantly represented in the corpus.

At the orthographic level, this corpus contains about 26,000 word tokens, corresponding to 6,000 different word forms. These word occurrences can be further subclassified into roughly 1,500 proper names (corresponding to 1,000 different word forms), 600 numerals (200 word forms), and 200 acronyms and abbreviations (90 word forms), the remaining lot being composed of common words.

The specificities of the corpus in terms of its vocabulary were calculated with respect to the 1950-1990 period of the "Trésor de la langue française" corpus (Imbs, 1971). This study revealed no marked deviation in terms of vocabulary.

Manually transcribed, the reference corpus contains a grand total of 85,000 phonemic symbols. Based on indications provided by the corpus producer, one can estimate the number of possible cases of liaisons to be approximatively 1,500, amongst which about 600 are compulsory. Similarly, the transcribed corpus contains 8,500 cases of mute-e, further subdivised between 1,000 obligatory deletions, 1,500 obligatory realisations and 6,000 optional deletions.

4. Results

Once the corpus selected, it was possible to have it transcribed by each participant, to compare the transcriptions and score the systems. A first computation of results was released; then, an adjudication stage gave the participating teams the opportunity to contest some of their errors. A new version of the corpus was then produced, and new scores were accordingly computed. Overall, the eight systems managed to fare relatively well with the task at hand, since they achieved at least 97% phonemes correct. The raw results, obtained at the term of this adjudication phase, are displayed in Table 1.

This table distinguishes between correctness and accurracy: the former gives the percentage of phonemes correctly predicted, whereas the latter also takes into account the percentage of spurious insertions. It is also to note that systems significantly differ in their treatment of optional phonemes: this fact is reflected in the important variability (nearly 5%) in the total number of phonemes produced.

But what is really needed is a classification of errors enabling us to pinpoint the most problematic cases for each system. The participating teams were consequently asked to work on a manual classification of their errors. In order to make these detailed results comparable, a common grid was defined, along the four following dimensions:

- related orthographical form;
- major grammatical category - in fact, we only distinguished between proper names, acronyms, number and symbols, and lexical items) -;
- error type (e.g. heterophonous homograph, morphological ambiguity, lean word, obligatory or forbidden liaison, schwa problem, preprocessing);
- typographical characteristics (e.g. upper case, digit, Roman figure).

This annotation scheme is unsatisfactory in many respects: f or instance, an incorrect preprocessing may result in a liaison error; a morphological boundary may be missed in a

borrowed word, etc. Nonetheless, this categorisation, together with the annotated mistakes provided by six of the 8 participating teams enabled us to conduct a more detailed quantitative analysis of the errors. The figures presented in Tables 2 and 3 cannot be taken at face value, but merely reflect the relative importance of various GP conversion problems in French, and the current ability of our systems to cope with those. It should finally be noted that, these (indicative) figures refer to occurrences of erroneous words.

Table 2 presents the distribution of errors by word category for 6 systems. It mainly illustrates the difficulties of correctly pronouncing proper names and acronyms. While these token only represent respectively 5.8% and 0.7% of the words in the test corpus, they are more strongly represented in the erroneous words. Errors on proper names represent between a half and a sixth of the total number of errors.

Numbers and numerals are another significant cause of errors, which can be further subclassified into 3 main categories:

- deletion of the final consonant of *cinq (5), six (6), huit (8), dix (10)* before a consonant, within dates, addresses or phone numbers;
- insertion of a segment corresponding to forbidden liaisons;
- substitution, especially due to the pronunciation *un* (instead of *une*) for *1* before a feminine noun.

Table 3 displays the distribution of errors by error type. It may be analysed by line or by column. It can also be analysed by line or by column. These breakdoowns show that the results are quite different from a system to another: the best system is not necessaril the best for coping with each difficulty. Though, it appears that, on the whole, the main problems are (foreign) proper names and the schwa. The latter point needs to be moderated though: the transcribed corpus contains 8,500 schwas, a small proportion of which (approximately 2% on the average) is in fact erroneously predicted.

5. Discussion

In this section, we critically review the methodology we used, and put forward several arguments and questions, to point out possible improvements. This part is not very much language dependent: the same types of methodological problems would certainly be encountered is other langages as well.

The definition of a common phonemic alphabet and of a universal grid for evaluation proved to be difficult and controversial, which is to be linked with the limits of a modular approach of assessment. An alternative solution to the one we adopted, to specify the inventory of phonemes and to describe the variability is to consider phonetic transcription not as a succession of atomic symbols, but as a succession of more complex units. Let us examine how this would work in the case of liaisons. Using abstract symbols, it is possible to represent optional liaisons in /z/ with the capital letter /Z/, which covers /z/ and zero in its realisation (equivalent classes reminiscent of the concept of archiphonemes). As a consequence, the two variants need not be explicitely listed in the test corpus. The use of this kind of alphabet allows us to evaluate not only the *accuracy* of a given GP conversion device, but also its *precision*. Let us assume, for instance, that the reference pronunciation of the first *o* in *microcosme* can be either /o/ or /ɔ/. In this situation, a system capable of predicting both timbres should be more rewarded than a system which only predicts one of the two possible outcomes, since arguably, the former is more precise than the latter. However, this kind of encoding could only capture some cases of phonological variability, namely the cases where the pronunciation of one single symbol is subject to variation, irrespective of its phonological context. More complex cases, where the alternative exists between several sequences of phonemes, or where there is a contextual dependency between adjacent symbols, would still need be explicitely enumerated - such cases are not uncommon. In addition, it is possible that the evaluation of the precision of a GP conversion device is not central in the specific context of speech synthesis systems, and that its undertaking would unduly complexify the scoring measure.

Another point is worth mentioning, which concerns a dimension along which systems were not evaluated. The reference corpus contains a lot of independently specified variants. A consequence is that this scoring strategy fails to assess the stylistic coherence of a given transcription. This has been found to be quite a minor problem: all the tested systems use deterministic transcription rules, which makes them unable to produce a kind of “incoherent” output.

It is however possible to enrich the test corpus annotation scheme, including for instance part-of-speech tags and an alignement between orthographic and phonemic strings at the word level; and a specification of the places where liaison is not possible.

Several test corpora could also be generated with the same text material. In this way, a system’s behaviour could also be tested by using different variants, each representing a different level of difficulty for the GP conversion task. Going from the most complex to the easiest, the first variant would be pre-segmented, the next would be typo-free, the next would further contain expanded abbreviations, an even easier would include tags (or even brackets) etc. Given the availability of accurate natural language processing tools for performing these tasks, these variants could be produced nearly automatically. An alternative exists regarding the constitution of a reference transcription for a text corpus: ear-transcribed or hand-transcribed, which means observed or not. The comparison of reference dictionaries (Juillard, 1965; Warnant, 1987; Larousse, 1989; Robert, 1990; Boë & Tubach, 1992) reveals the obvious difference of phonemic prediction by the authors. More generally, it seems that the problem of finding an agreement between different listeners makes the constitution of ear-transcribed corpora difficult. It thus appeared more reasonable to opt for a hand-transcribed cor-

pus in these experiments. This choice was also plainly justified with respect to the participating GP systems themselves, whose pronunciation rules reflected more the content of pronunciation dictionaries than the actual pronunciations.

Common word lexica are very valuable tools for measuring the capabilities of a transcription system to handle phenomena internal to lexemes or derivational morphemes. Moreover, their coverage enables us to test the transcription capabilities on a large-scale, i.e. to evaluate overall consistency of the transcription rules. Finally, specialized lexica, such as proper names lexica, constitute suitable testing conditions for evaluating GP conversion devices on a task like reverse directory inquiry. Nevertheless, irrespective of the public availability of large scale phonetic dictionaries for French, testing GP conversion on systems on running texts has the advantage, in comparison to allow us to evaluate the ability of a device to properly process contextual variants (final schwas, liaisons, etc.).

6. Conclusion

The evaluation results presented in this paper give an accurate picture of the state-of-the-art in the domain of GP conversion for French, and of the problems still to be solved: proper names, acronyms, numbers or symbols, loan words, heterophonous homographs, preprocessing, schwa and liaison. Moreover, the ressources and data will be put at the disposal of the scientific and industrial community, in order to serve as reference for future benchmarks. This will enable a quantitative analysis of the results obtained, and a measurement of the progress achieved for each specific system. For this goal, a corpus was defined, based on newspaper texts. A working methodology was designed, and tests were performed, using 8 systems.

The discussion on methodology raised serious theoretical objections against the concept of an objective evaluation of GP conversion devices. However, it must be emphasised that many methodological issues are still difficult to handle. We hope that this aspect of the paper, namely the discussion on pros and cons of the methodology, will help evaluations of similar systems for other languages.

Acknowledgments

The experiments reported on in this paper were conducted in the framework of the ILOR project supported by the Francil network. This project was funded by the AUPELF-UREF (Association of French-speaking universities). We are grateful to the French newspaper *Le Monde* who kindly provided us with electronic text material. Special thanks to Pascale Bouchet and Roseline Hamamdjan for their work on the corpus transcription.

References

[d'Alessandro, C., Aubergé, V., Béchet, F., Boula de Mareüil, P.] Foukia, S., Goldman, J.-P., Isabelle, J.-F., Keller, E., Marchal, A., Mertens, P., Pagel, V., O'Shaughnessy, D., Richard, G., Talon, M.-H., Wehrli, E., Yvon, F. (1997). "Vers l'évaluation de systèmes de synthèse de parole à partir du texte en français". *Premières Journées Scientifiques et Techniques du réseau Francil de l'AUPELF-UREF*, pp. 393-397. Avignon.

[Aubergé, V. (1991)] *La synthèse de la parole : "des règles aux lexiques"*. Doctoral Dissertation. Université P. Menès-France, Grenoble.

[Béchet & El-Bèze, M. (1996)] "Intégration de différents niveaux linguistiques pour le traitement des mots hors-dictionnaire dans la conversion graphème-phonème automatique". *XXI^{es} Journées d'Étude sur la Parole*, pp. 421-42. Avignon.

[Benoît, C. (1997)] "Evaluation inside or assessment outside? ". In *Progress in Speech Synthesis* (van Santen, J.P.H., Sproat, R.W., Olive J.P., Hirschberg, J. Eds). Springer-Verlag, New York.

[Boë, L.-J. & Tubach, J.-P. (1992)] *De A à Zut. Dictionnaire phonétique du français parlé*. ELLUG, Grenoble.

[Boula de Mareüil, P. (1995)] "Vers la phonématisation automatique des sigles", *La linguistique*, 31:1, pp. 93-103.

[Boula de Mareüil, P. (1997)] "Conversion graphème-phonème : de la formalisation à l'évaluation". *Premières Journées Scientifiques et Techniques du réseau Francil de l'AUPELF-UREF*, pp. 399-406. Avignon.

[Carlson, R., Granström, B. & Nord, L. (1990)] "Segmental evaluation using the Esprit/SAM test procedures and monosyllabic words". In *Talking Machines* (Bailly, G. & Benoit C. eds), pp 443-453.

[Dutoit, T. (1993)] *High-quality text-to-speech synthesis of the French language*. Doctoral Dissertation. Faculté Polytechnique de Mons.

[Gibbon, D., Moore, R. & Winski, R. eds, (1997)] *Handbook of Standards and Ressources for Spoken Language Systems*. Mouton de Gruyter, Berlin.

[Gaudinat, A. & Wherli, E. (1997)] "Analyse syntaxique et synthèse de la parole: le projet FIPSVox". *Traitement Automatique des Langues* 38:1, pp. 121-134.

[Imbs, P. (1971)] *Trésor de la Langue Française. Dictionnaire de la langue du XIX^{e} et du XX^{e} siècles (1989-1960)*. Éditions du CNRS, Paris.

[Juilland, A. (1965)] *Dictionnaire inverse de la langue française*. Mouton & Co., London-The Hague-Paris.

[Keller, E. (1997)] "Simplification of TTS architecture vs. operational quality". *Eurospeech*. Rhodes.

[Klaus, H., Fellbaum, K. & Sotscheck, J. (1997)] "Auditive Bestimmung und Vergleich der Sprachqualität von Sprachsynthesesystemen für die deutsche Sprache". *Acta Acustica* 83, 124-1.

[Kraft, W. & Portele, T. (1995)] "Quality evaluation of five German speech synthesis systems". *Acta Acustica* 3:4, pp. 351-365.

[Larousse, P. (1989)] *Dictionnaire de la langue française lexis*, Larousse, Paris.

[Martinet, A. (1960)] *Éléments de linguistique générale*. Armand Colin, Paris.

[O'Shaughnessy, D. (1984)] "Design of a Real-Time French Text-to-Speech System". *Speech Communication* 3:4, pp. 317-324.

[Pols, L.C.W. & Jekosch, U. (1997)] "A structured way of looking at the performance of text-to-speech systems". In *Progress in Speech Synthesis* (van Santen, J.P.H., Sproat, R.W., Olive J.P. & Hirschberg, J. eds), Springer-Verlag, New York.

[Robert, P. (1990)] *Dictionnaire alphabétique & analogique de la langue française*. Société du Nouveau Littré, Paris.

[Silverman, K., Basson, S. & Levas, S. (1990)] "Evaluating synthesiser performance: is segmental intelligibility enough? ". *ICSLP 90*, pp. 981-984, Kobe, Japan.

[Sorin, C., Emerard, F. (1996)] "Domaines d'application et évaluation de la synthèse de parole à partir du texte". In *Fondements et perspectives en traitement automatique de la parole* (Méloni, H. ed), pp. 123-131. AUPELF-UREF.

[van Bezooijen, R. & Pols, L.C.W (1990)] "Evaluation of text-to-speech systems: Some methodological aspects". *Speech Communication* 9, 263-270.

[van Santen, J.P.H. (1993)] "Perceptual experiments for diagnostic testing of text-to-speech systems". *Computer Speech and Language* 7, 49-100.

[Warnant, L. (1987)] *Dictionnaire de la prononciation française dans sa norme actuelle*. Éditions Duculot, Paris-Gembloux.

[Yvon, F. (1996)] *Prononcer par analogie : motivations, formalisation et évaluation*. Doctoral Dissertation, ENST, Paris.

Labs	#Phonemes	Correctness	Accuracy
Lab. A	83841	97.1	93.0
Lab. B	84250	94.9 (*)	94.4 (*)
Lab. C	85850	97.7	97.3
Lab. D	85554	98.4	97.8
Lab. E	86338	99.2 (*)	98.8 (*)
Lab. F	86205	99.2	98.7
Lab. G	86938	99.3	99.0
Lab. H	86047	99.6	99.5

TAB. 1 - *Global performance of the eight systems*

For each system, this table gives successively the total number of phonemes predicted, the system's correctness and accuracy.

(*) These figures largely underestimates this system's performance, which have been severly degraded in terms of phonemic correctness and accuracy by a non-negligible number of entirely incorrect sentences. A further exam of these sentences revealed that they either were incorrectly formatted, or had been wrongly aligned with the reference corpus.

Labs	Lab. B	Lab. C	Lab. D	Lab. F	Lab. G	Lab. H
Proper name	180 (9.6)	296 (24.0)	204 (22.7)	113 (15.2)	168 (25.6)	131 (49.4)
Acronym	105 (5.6)	30 (2.4)	73 (8.1)	15 (2.0)	11 (1.7)	7 (2.6)
Number or symbol	113 (6.0)	157 (12.7)	55 (6.1)	90 (12.1)	50 (7.6)	30 (11.3)
Other	1469 (78.7)	752 (60.9)	565 (63.0)	526 (70.7)	428 (65.1)	97 (36.6)
Total	1867	1235	852	744	657	265

TAB. 2 - *Distribution of errors by word category for 6 systems*

For a given system and word category, each cell of this table contains the absolute number of errors for this category, and the corresponding percentage of the errors for the system.

Labs	Lab. B	Lab. C	Lab. D	Lab. F	Lab. G	Lab. H
Lean word	77 (4.1)	52 (4.2)	129 (15.1)	97 (13.0)	89 (13.5)	103 (38.9)
Liaison	111 (5.9)	123 (10.0)	76 (8.9)	49 (6.6)	38 (5.8)	15 (5.7)
Schwa	493 (26.4)	374 (30.3)	134 (15.7)	46 (6.2)	91 (13.9)	3 (1.1)
Heterophonous homograph	98 (5.2)	47 (3.8)	34 (4.0)	90 (12.1)	7 (1.1)	19 (7.2)
Typing errors	116 (6.2)	78 (6.3)	14 (1.6)	15 (20.1)	38 (5.7)	15 (5.7)
Preprocessing	537 (28.8)	187 (15.1)	79 (9.3)	19 (2.6)	26 (4.3)	53 (20.0)
Morphological ambiguity	197 (10.6)	1 (0.1)	15 (1.8)	8 (1.1)	2 (0.3)	6 (2.3)
Other	238 (12.7)	373 (30.2)	371 (43.5)	420 (56.4)	366 (55.7)	51 (19.2)

TAB. 3 - *Distribution of errors by type for 6 systems*

For a given system and error type, each cell of this table contains the absolute number of errors for this type, and the corresponding percentage of the errors for the system.