# A joint intelligibility evaluation of French text-to-speech synthesis systems: the EvaSy SUS/ACR campaign

**Philippe Boula de Mareüil[1], Christophe d'Alessandro[1], Alexander Raake[1], Gérard Bailly[2], Marie-Neige Garcia[3], Michel Morel[4]**

[1] LIMSI-CNRS, BP 133 — F-91403 Orsay CEDEX, { mareuil;cda;Alexander.Raake}@limsi.fr
[2] ICP, 46 avenue F. Vallet, F-3%8031, Grenoble, bailly@icp.inpg.fr
[3] ELDA, 55–57 rue Brillat Savarin, 75013 Paris, garcia@elda.org
[4] CRISCO, Université de Caen , 14032 Caen CEDEX, morel@crisco.unicaen.fr

**Abstract**

The EVALDA/EvaSy project is dedicated to the evaluation of text-to-speech synthesis systems for the French language. It is subdivided into four components: evaluation of the grapheme-to-phoneme conversion module (Boula de Mareüil *et al*., 2005), evaluation of prosody (Garcia *et al*., 2006), evaluation of intelligibility, and global evaluation of the quality of the synthesised speech. This paper reports on the key results of the intelligibility and global evaluation of the synthesised speech. It focuses on intelligibility, assessed on the basis of semantically unpredictable sentences, but a comparison with absolute category rating in terms of e.g. pleasantness and naturalness is also provided. Three diphone systems and three selection systems have been evaluated. It turns out that the most intelligible system (diphone-based) is far from being the one which obtains the best mean opinion score.

## 1. Introduction

Evaluation is of a crucial concern for linguistic engineering, in terms of adequacy to market needs, diagnostics and computational performance. In this article we present some of the results of a project devoted to the diagnostic evaluation of language and speech processing devices for the French language, including text-to-speech (TTS) synthesis. The EVALDA/EvaSy project is intended to expand upon the ARC AUPELF (now AUF) campaign of 1996–1999, the only previous formal evaluation campaign for TTS systems in French.

The aim of the EvaSy project is to assess both diphone-based TTS systems and the new generation of speech synthesis systems. The latter systems rely on large corpora consisting of careful recordings of trained voice talents, as well as selection and concatenation algorithms. Although these systems appear very fluent and naturally sounding, their intelligibility has never been formally rated to the best of our knowledge (at least in French). However, in some applications such as reading machines for the blind, intelligibility (especially at high speech rates) may be more important than naturalness. It is thus relevant to know whether these systems perform better than older diphone-based systems in terms of intelligibility. The originality of this paper is to draw a parallel between "good" and "beautiful" systems.

The intelligibility test we designed is based on semantically unpredictable sentences (SUS), a paradigm that allows an objective assessment of word-level intelligibility. Six state-of-the-art TTS systems for French were tested in the SUS campaign: three were diphone-based systems (referred to as D1, D2, D3) and three were selection/concatenation systems (referred to as S1, S2, S3). Although the detailed results are kept anonymous, we are in a position to say that they came from CRISCO, ICP, LIMSI-CNRS, Multitel, Elan and Babel (now Acapela group). References to these systems can be found in d'Alessandro & Tzoukermann (2001).

A list of semantically unpredictable sentences was built — based on 4 of the 5 syntactic structures initially proposed by Benoît (1990). The design and optimisation of the underlying lexicon and speech material are described in more detail in another communication presented at this conference (Raake & Katz, 2006). Preliminary tests using natural speech and noisy speech demonstrated the robustness of the test protocol and platform.

The following sections describe the corpus and protocol used for a direct measurement of intelligibility. In Section 5 and in the conclusion, results are also compared to those of an absolute category rating test including mean opinion score and comprehension (distinguished from intelligibility).

## 2. Corpus

A list of 288 semantically unpredictable sentences was built, divided in blocks including 4 syntactic structures:
(1) adverb det. $Noun_1$ Verb-*t*-pron. det. $Noun_2$ Adjective ?
(2) determiner $Noun_1$ Adjective Verb determiner $Noun_2$
(3) det. $Noun_1$ $Verb_1$ determiner $Noun_2$ *qui* ("that") $Verb_2$
(4) determiner $Noun_1$ Verb preposition determiner $Noun_2$

Structure 3 originally proposed by Benoît (1990) was not kept, because it only contained 3 target words (nouns, verbs or adjectives, here written with a capital initial letter) instead of 4 in the other structures. Each block was composed of 12 sentences, as in the following sample (Table 1):

In order to have comparable sentences and blocks, all content words were singular, monosyllabic (unless a final schwa was uttered) and had a high frequency of use according to the BRULEX lexicon (Content *et al*., 1994). Prepositions were also monosyllabic: e.g. *sur* ("on"). Determiners were definite articles ("the") *le, la* or *l'* before a vowel. Adjectives which are normally located before nouns in French or which were homonyms of verb forms were excluded. For the remaining ones, the agreement in gender with nouns and determiners was checked. In each sentence, the first noun (which also had to agree grammatically with the anaphoric pronoun in the first construction) was different from the second noun. Verbs, in the third person present tense, had to be transitive in structures 1 and 2 (as $Verb_1$ in structure 3)

and could be intransitive in structure 4 (as $Verb_2$ in structure 3). Whether they were transitive or intransitive, the possibility that they could be used with no complement was carefully watched: e.g. *songe* ("thinks"). Finally, some tokens which might have raised pronunciation issues (such as heterophonous homographs) were discarded.

| | |
|---|---|
| La loi brille par la chance creuse. | (4) |
| La classe gaie montre le frein. | (2) |
| Quand le lien signe-t-il l'onde pleine ? | (1) |
| Le test clair mange la haine. | (2) |
| L'or jaune porte le dôme. | (2) |
| Comment la soif lance-t-elle le bol proche ? | (1) |
| Le mur siffle la buée qui vole. | (3) |
| La banque dit la dinde qui plaît. | (3) |
| La terre dresse la boîte qui rage. | (3) |
| Où l'oeuf cite-t-il le thé doué ? | (1) |
| Le nom luit sur le bras nu. | (4) |
| Le choix tape dans la queue close. | (4) |

Table 1: sample of semantically unpredictable sentences with the type of their syntactic structure.

Once the word material was designed (over 400 target words), SUS lists were randomly generated. Several trials were made, and the one which provided the most balanced distribution of phonemes across the blocks was retained. The phoneme repartition by block was compared to that of two authoritative French lexica — BRULEX (Content *et al*., 1994) and LEXIQUE (New *et al.*, 2004) — according to chi-square tests. The resulting list was tuned manually, and some words were exchanged within a same block if by chance an automatically yielded word sequence made sense. This way, no block was favoured: configurations which could have helped understanding were avoided. The final SUS list was definitely meaningless, it was more thoroughly controlled than those of previous studies, and can be used as a reference for various experiments.

## 3. Protocol

The SUS list was read by a professional male speaker in a soundproof booth, and the recordings were sampled at at 16 kHz (16 bits, mono) in the Wave format. In addition to this natural reference, 6 systems were tested. The participating teams had to synthesise the 288 sentences mentioned above within a few hours, at the same sampling rate as the natural reference. For the test strictly speaking, the organiser (the European agency ELDA) retained 3 blocks of 12 sentences for any of the natural or synthetic voices (without counting one sentence per voice for a familiarisation phase). This way, 22 of the 24 blocks were used. For each system or voice the blocks were different, but since they were designed to be comparable, the bias was believed to be minor. All the sentences (whose sound level was equalised beforehand) were presented only once, in a random order that was different for each listener. The test lasted between 2 and 3 hours per subject.

The test protocol was automated: an interface was designed to capture the subjects' responses and to analyse the results. Listeners were asked to orthographically transcribe the sentences they heard. The typed sentences were then transcribed phonetically with the help of a grapheme-to-phoneme converter whose word error rate was less than 1 % (Boula de Mareüil, 1997) and compared to the phonetic transcription of the reference sentences, in order not to count homophonous responses and spelling mistakes as errors. In semantically unpredictable sentences for instance, *voix* ("voice") and *voie* ("way") are equally correct since they read the same (/vwa/).

The SUS test campaign was conducted at ELDA (Paris), with 19 listeners in a quiet environment, using high-quality audio material and an on-line evaluation platform. The subjects were 19–46 year-old native French speakers with no known hearing problem. They were not experts in speech synthesis; they were paid for this task.

## 4. Results

A first way to rank the systems consists in considering any sentence that is not phonetically identical to the original as erroneous, following Benoît *et al*. (1996). Second, we counted the target words that were not properly reproduced, for a more fine-grained evaluation. The scoring was based on the `sclite` dynamic programming algorithm (http://ww.nist.gov/speech/tools/). The scores were restricted to the 4 SUS target words, whose phonetic transcriptions were also split into phonemes so as to go further in the analysis and measure phoneme accuracy rates. With this goal in view, listeners' transcriptions which contained more than 4 target words (after discarding determiners, adverbs, pronouns and prepositions) were carefully checked. Due to spelling or typographic errors, they represented about 5% of the sentences.

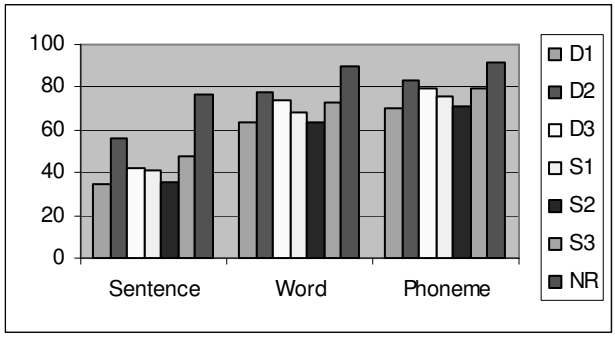


Figure 1: percent correct transcriptions in the SUS test on sentences, target words and phonemes for the 6 systems and the natural reference (NR).

As is apparent in Figure 1, results correlate whether sentences, words or phonemes are considered. On sentences, words or phonemes, the best percentages are achieved by the natural reference, D2, S3, D3, S1, S2 and D1 in a decreasing order. Diphone systems, based on hyper-articulated units, are rather more intelligible than are selection systems (45% vs. 41% on sentences, on an average), and all of them are far from a natural voice. What is more, the percentages of correct sentences reported by Benoît (1990), ranging from 28.6% and 58.1%, are close to ours. This suggests that intelligibility has little improved for more than 15 years.

The percentages of insertions, deletions and substitutions have also been examined with the help of the `sclite` software. Words such as *toit* (/twa/, "roof") and *doigt* (/dwa/, "finger") for instance, with a /t/~/d/ confusion, are among the most frequently misunderstood

words. Whether on words or on phonemes and whatever the system or voice, the highest rates are deletion rates. They reflect the fact that when listeners did not understand a word at all, they resorted to question marks or suspension points (which were of course filtered out in the computation of the results). They therefore represent a more severe criterion than substitution rates, and the wider ranges of results they exhibit enable a more clear-cut discrimination between the different systems. The results reported in Table 2 preserve the hierarchy between the systems, and support those of Figure 1.

| %deletion | D1 | D2 | D3 | S1 | S2 | S3 | NR |
|---|---|---|---|---|---|---|---|
| Word | 25.4 | 15.5 | 18.5 | 22.4 | 24.7 | 15.9 | 7.1 |
| Phoneme | 26.5 | 14.6 | 18.6 | 21.4 | 25 | 17.1 | 7.2 |

Table 2: percentages of word and phoneme deletions in the SUS test (NR = natural reference).

## 5. Comparison with an ACR test

### 5.1. Procedure

The SUS test was complemented by a series of speech quality tests employing the classical 5-point absolute category rating (ACR) scale often used in telecommunications + 1 point to each extremity of the scale to avoid the saturation effect (Möller, 2000). In addition to mean opinion score (MOS), six categories have been retained: comprehension, pleasantness, non-monotony, naturalness, fluidity and pronunciation. They were adapted to the French language from the criteria proposed for earlier speech synthesis tests (d'Alessandro, 2004), especially the Vermobil project (Kraft & Portele, 1995). Table 3 displays how the listeners were requested to respond. The scale that was presented to the subjects was continuous, and larger than the 5 points which were glossed as in Table 3, in order to encourage the use of the extreme points.

As above, the participants were asked to synthesise hundreds of sentences within a short lapse of time. Here, the corpus was EUROM 1 (Campione & Véronis, 1998), developed and collected within the framework of the MULTEXT (Multilingual Text Tools and Corpora) and Esprit 2589/SAM (Multilingual Speech Input/Output Assessment Methodology and Standardisation) projects. For each language including French, the database is made up of 40 passages of about 20 seconds: 5 sentences linked by a coherent thematic structure. The EUROM 1 recordings, which were made in an anechoic room, are of a good acoustic quality. Ten speakers (5 males, 5 females) read 15 passages on an average, resulting in a 1-hour corpus.

We selected a subset of 20 passages, on which subjective tests were carried out. Each paragraph was read by a natural voice and 6 systems (in the Wave format, beforehand equalised, at a 16 kHz sampling rate, 16 bits, mono), in a random order that was different for each trial. The test lasted 3 hours with breaks every 20 minutes. It also took place at ELDA, throughout a user-friendly interface which allowed the speech samples to be appreciated. It was administered to 17 subjects whose profile was the same as in the SUS test.

| **MOS (very bad — very good)** |
|---|
| Comment appréciez-vous globalement ce que vous venez d'entendre ? Très mauvais — très bon |
| **Comprehension (very difficult — very easy)** |
| Comment décririez-vous la facilité à comprendre le message ? Très difficile — très facile |
| **Pleasantness (very unpleasant — very pleasant)** |
| Comment décririez-vous cette voix ? Très désagréable — très agréable |
| **Non-monotony (very monotonous — very varied)** |
| Évaluez le caractère monotone ou varié de ce que vous venez d'entendre : très monotone — très varié |
| **Naturalness (very artificial — very natural)** |
| Comment apprécieriez-vous le naturel de ce que vous venez d'entendre ? Très artificiel — très naturel |
| **Fluidity (very jerky — very fluid)** |
| Comment appréciez-vous le côté haché ou fluide de l'élocution ? Très haché — très fluide |
| **Pronunciation (serious problems — no problem)** |
| Avez-vous remarqué des problèmes de prononciation ? Très gênant — aucun problème |

Table 3: questions asked to the subjects in the ACR test, and paraphrase of the corresponding extreme category responses (in French).

### 5.2. Results

From the results (Figure 2), it appears that in all the categories the natural reference obtains the highest scores (above 4), before S3 and next D3. S3 crosses the 4/5 threshold for comprehension and pronunciation and is the only system that crosses the 3/5 threshold in the other categories: it is by far the best system, one point on average above the second system, D3. D3 is the only system that is rated between 2 and 4 in all the categories (even naturalness). That is, unit selection systems are not necessarily judged better than are diphone systems. Interestingly, D1 and S1 on the one hand, D2 and S2 on the other hand are vying with each other as far as MOS and naturalness are concerned. The worst score is assigned for the latter category to D1. However S1 and S2 are the worst systems respectively for pleasantness and pronunciation. D2 is more understandable (but also more monotonous) than both S1 and S2.

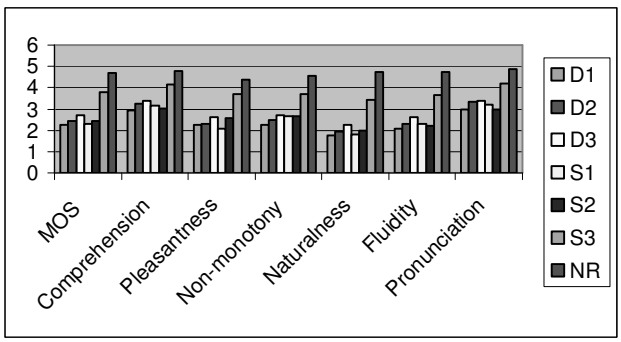


Figure 2: results of the ACR test for the 6 systems and the natural reference (NR). Scores are mapped to a 1–5 scale, 5 being the highest mark.

## 6. Conclusion and future work

In summary, we developed a comparative approach to TTS evaluation: we designed a SUS test and a more impressionistic ACR test which among other things allowed us (in other researchers' wake) to distinguish intelligibility from comprehension. A diphone system (D2) was awarded as the most intelligible in the SUS test, but a selection system (S3) was very much preferred in the ACR test. Also, even though our results are dependent upon the assessed systems, both tests highlight the fact that selection systems are not necessarily better than are diphone systems. This holds true for the prosody evaluation (Garcia *et al*., 2006).

Statistical analyses examining the significance of the results as well as the subject and stimulus effects are required for future investigation, in particular to pinpoint systems showing high or low correlations between intelligibility and quality measures. Second, an acoustic analysis of the most error-prone segments in the intelligibility test would be of major interest. Based on these analyses, implications for the usage of the systems for different types of applications can be pointed out. For example, a highly intelligible system, in spite of a poor voice quality, may be a good choice for applications like reading machines for visually-impaired people. In return, TTS systems suited to be employed at a large scale in spoken dialogue systems may show a higher need for voice quality, since they can be viewed as the business card of the system; lower intelligibility values may be less problematic for such systems. Enormous efforts have been undertaken and considerable means have been involved to shift from an announcement style to more intimate and expressive situations, with conversational speech mannerisms such as laughter (Campbell *et al.*, 2005). This is a stimulating challenge, but before considering a paradigm change from reading machines to talking machines (and regardless of funding issues), users' profile and environment should be taken into account to avoid an application/evaluation divorce.

A by-product of this research is a series of carefully controlled SUS lists, a valuable and reproducible resource for intelligibility testing. In order to find a sensitive but efficient intelligibility measure capable of distinguishing the different systems more clearly, an alternative regarding the test design will be addressed. The list-wise repartition of sentences enables an adaptive speech reception threshold measurement against noise:

- targeting 50% correct keywords
- using one of two different types of stationary noise maskers (one masker signal matching the overall long-term spectrum of all synthesis systems, or different maskers for the different systems, each matching the long-term spectrum of that system).

Several intelligibility tests can be conducted using this method.

## 7. Acknowledgements

The EVALDA evaluation campaign is financed by the French Ministry of Research in the context of the Technolangue programme. We are also grateful to Julien Coucoureux and Damien Damour for their help in designing the test interface and analysing the results.

## 8. References

d'Alessandro, C. (2004), L'évaluation des système de synthèse la parole. In S. Chaudiron, (ed.), *Evaluation des systèmes de traitement de l'information*, Hermès, Paris, pp. 215–239.

d'Alessandro, C. & Tzoukermann, E. (eds) (2001), *Synthèse de la parole à partir du texte, Traitement Automatique des Langues,* 42(1), Hermès, Paris.

Benoît, C. (1990), An intelligibility test using semantically unpredictable sentences: towards the quantification of linguistic complexity, *Speech Communication*, **9**(4), pp. 293–304.

Benoît, C., Grice, M., Hazan, V. (1996), The SUS test: a method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. *Speech Communication*, **18**(4), pp. 381–392.

Boula de Mareüil, P. (1997), *Étude linguistique appliquée à la synthèse de la parole à partir du texte*, PhD thesis, University of Paris XI, Orsay.

Boula de Mareüil, P. d'Alessandro, C., Bailly, G., Béchet, F., Garcia, M.-N., Morel, M., Prudon, R., Véronis, J. (2005). Evaluating the pronuncation of proper names by four French grapheme-to-phoneme convverters. In *Proceedings of the Ninth European Conference on Speech Communication and Technology (Interspeech)*, Lisbon, pp. 1521–1524.

Campbell, N., Kashioka, H., Ohara, R. (2005), No Laughing Matter. In *Proceedings of the Ninth European Conference on Speech Communication and Technology (Interspeech)*, Lisbon, pp. 465–468.

Campione, E. & Véronis, J. (1998). A multilingual prosodic database. In *Proceedings of the Fifth International Conference on Spoken Language Processing,* Sydney, 3163–3166.

Content, A., Mousty, P., Radeau, M. (1990), BRULEX : Une base de données lexicales informatisée pour le Français écrit et parlé. *L'Année Psychologique*, **90**, pp. 551–566.

Garcia, M.-N., d'Alessandro, C., Bailly, G., Boula de Mareüil, P. Morel, M. (2006), A joint prosody evaluation of French text-to-speech systems: the EvaSy Prosody campaign. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genoa.

Kraft, V. & Portele, T. (1995), Quality Evaluation of Five German Speech Synthesis Systems, *Acta acustica*, **3**, pp. 351–365.

Möller, S. (2000), *Assessment and Prediction of Speech Quality in Telecommunications*, Kluwer Academic Publishers, Boston.

New, B., Pallier, C., Brysbaert, M., Ferrand, L. (2004), Lexique 2: A New French Lexical Database. *Behavior Research Methods, Instruments, & Computers*, **36**(4), pp. 516–524.

Raake, A. & Katz, B.FG (2006), SUS-based Method for Speech Reception Threshold Measurement in French. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genoa.