A joint prosody evaluation of French text-to-speech synthesis systems

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

[1] ELDA, 55–57 rue Brillat Savarin, 75013 Paris, garcia@elda.org
[2] LIMSI-CNRS, BP 133 — F-91403 Orsay CEDEX, { mareuil;cda}@limsi.fr
[3] ICP, 46 avenue F. Vallet, F-388031, Grenoble, bailly@icp.inpg.fr
[4] CRISCO, Université de Caen , 14032 Caen CEDEX, morel@crisco.unicaen.fr

Abstract

This paper reports on prosodic evaluation in the framework of the EVALDA/EvaSy project for text-to-speech (TTS) evaluation for the French language. Prosody is evaluated using a prosodic transplantation paradigm. Intonation contours generated by the synthesis systems are transplanted on a common segmental content. Both diphone based synthesis and natural speech are used. Five TTS systems are tested along with natural voice. The test is a paired preference test (with 19 subjects), using 7 sentences. The results indicate that natural speech obtains consistently the first rank (with an average preference rate of 80%) , followed by a selection based system (72%) and a diphone based system (58%). However, rather large variations in judgements are observed among subjects and sentences, and in some cases synthetic speech is preferred to natural speech. These results show the remarkable improvement achieved by the best selection based synthesis systems in terms of prosody. In this way; a new paradigm for evaluation of the prosodic component of TTS systems has been successfully demonstrated.

1. Introduction

The EVALDA/EvaSy project is dedicated to the evaluation of text-to-speech (TTS) synthesis systems for the French language. It 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 EvaSy project is subdivided into four components: evaluation of the grapheme-to-phoneme conversion module (Boula de Mareüil *et al*., 2005), evaluation of prosody, evaluation of intelligibility, and global evaluation of the quality of the synthesised speech (Boula de Mareüil *et al*., 2006). One of the aims of the project is to assess the quality of the new generation of text-to-speech systems, those referred to as selection and concatenation systems compared to diphone based systems.

The prosody generated by the text-to-speech systems should be as close as possible to the prosody of natural voices. This study is based on the approach proposed by Prudon *et al*. (2004). The intonation contours of the different speech samples are transplanted on a same segmental content. This method allows us to evaluate the prosody module of each system independently of the other modules (text processing module and acoustic synthesis module).

In this paper we report on the evaluation of the prosody generated by the speech synthesis systems participating in the EVALDA/Evasy campaign. Five state-of-the art systems for French (3 diphone systems, referred to as D1, D2 and D3, and 2 non-uniform unit selection systems, referred to as S1 and S2) are tested in this project. They were designed by Acapela Group (Mons, B), CRISCO (University of Caen, F), ELAN (now part of the Acapela Group), ICP (CNRS/University of Grenoble, F), LIMSI-CNRS (Orsay, F). Natural voice (referred to as NR) is used as a reference.

The following section describes the method and the experimental material used for prosodic evaluation. Results of the evaluation are presented in Section 3. This will lead us to some conclusions in Section 4.

2. Method

2.1. Prosody transplantation

Our aim is a specific evaluation of the synthetic speech prosody, independently of other aspects of TTS (like grapheme-to-phoneme conversion, speaker voice quality or segmental quality). This is performed by using prosodic transplantation. The intonation contours are mapped onto a common segmental content (either diphones or natural speech in the present experiments).

2.2. Speech material

The evaluation corpus is composed of seven phonetically balanced sentences extracted from the BREF corpus (Lamel *et al*, 1991).

The spoken sentences durations are between about 4 and 11 s. There is some variation in duration between the different systems.

Six versions of each sentence are available: 5 are produced by the TTS systems participating in the campaign, the last one is the read sentence of BREF. The intonation contours of each sentence are used as input for a common segmental content.

In order to evaluate the influence of the segmental content databases on the prosodic transplantation, two conditions are used: the first one is the MBROLA French diphone database (Dutoit *et al*., 1996) and the second one is the natural voice itself, modified with the help of a high quality pitch and duration modification algorithm. These two conditions will be referred to as the "diphone voice" and the "natural voice".

For some systems, the prosodic data are directly generated by the systems in the ".pho" format (the format used in MBROLA). Then they are applied on the MBROLA database and natural speech. For some systems, the prosodic data are not available in the ".pho" format. Then the prosodic data are obtained with the help of an alignment tool: MBROLign (Malfrère *et al*., 1997). For each synthesised sentence, MBROLign performs a phoneme alignment and extracts the prosodic data (according to the ".pho" format). These data are then

applied on the MBROLA database and natural speech. As some details in the phonemic content may differ between the systems, this has to be checked manually before performing the phoneme alignment.
Average pitch differs between systems. In order to avoid a bias because of a mismatch between the segmental and prosodic contents, average pitch of all the sentences is normalised.

2.3. Protocol

Paired preference tests were conducted in Paris, at ELDA (Evaluations and Language resources Distribution Agency). 19 subjects (10 females, 9 males) participated in the experiments. The subjects were 20-40 year old native French speakers with no known hearing problem. They were not experts in speech synthesis; they were paid for the task. Testing took place in a quiet environment through headphones, using high-quality audio material and a specially designed on-line evaluation platform.
For each sentence, a pair of stimuli were presented to the subjects (the same sentence with two different intonation contours), and they were asked to indicate which version they preferred. All possible sentence combinations were considered.
To avoid a learning effect, the comparison pairs were randomised and presented in different orders to different subjects. The subjects could listen to the sentences as many times as they wanted, but they were instructed to rather make a judgement on the basis of their first impression.
Subjective tests are performed under the MBROLA diphone and natural speech conditions, but the subjects were not informed of the underlying segments used.
Preliminary tests demonstrated the robustness of the test protocol and platform. In particular, we organised a successful comparison test in which the task was to compare the synthetic sentences before and after prosodic transplantation.
In summary, the variables of this test were: 1. the TTS systems; 2. the voices ("diphone" and "natural"); 3: the sentences; 4: the subjects. Of course the "TTS system" variable was the most interesting one for our purpose. Results are discussed in the next section.

3. Results

3.1. Global Preference rates

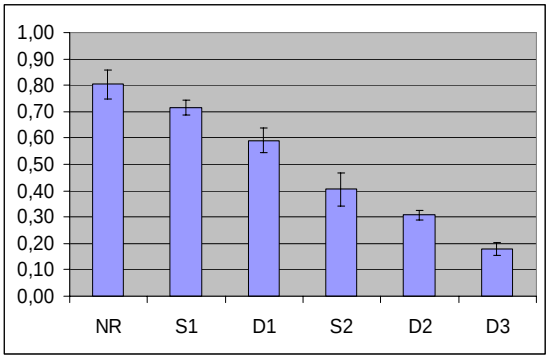

Figure 1: Global preference rates for each system. (Error marks are confidence intervals)

Global Preference rates for each system (on both "diphone voice" and "natural voice") are plotted in Figure 1. Global preference rates are computed as the number of times a system is preferred in a pair divided by the number of pairs. The number of stimuli pairs presented for each system is 665. The natural reference (NR) reaches a preference rate of 80%, followed by S1 (71%) and D1 (58%). Preferences rate below 50% are obtained for systems S2, D2 and D3 (respectively 40%, 31% and 18%).
Differences between systems are conspicuous. This global scoring indicates that S1 is clearly preferred. However, the second system in ranking is a diphone based system. Then selection systems are not necessarily better than diphone systems. It seems that fine tuning of the system is the key for high quality prosody. The prosodic ranking obtained by the best system is lower than the natural reference. This is in contrast with the results of Prudon *et al* (2004). In their study, the diphone system was judged even better than the natural reference. A possible explanation was the monotonous quality of the speaker who produced the natural reference.

3.2. Differences between voices

Ideally, the results should not be dependant on the segmental basis used for prosodic transplantation. In order to test this hypothesis, two different techniques are used. In the first technique we perform the prosodic transplantation on a widely used diphone synthesis system (MBROLA) whereas in the second technique a pitch and duration modification is directly applied on the natural sentence. Note that the same prosodic description file (the ".pho" format proposed in the MBROLA project) is used in both cases. Global preference rates for the "diphone voice" and "natural voice" are reported in Figure 2.
This Figure shows that the preference order of the systems is robust against the two different segmental bases.

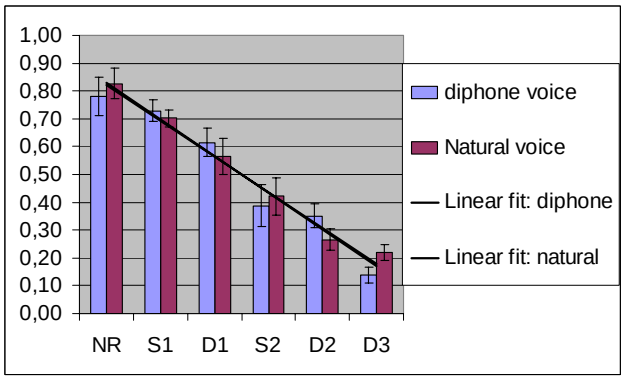

Figure 2: Global preference rates for each system, "natural voice" and "diphone voice" (Error marks are confidence intervals)

Differences between the mean preference rates for a same system are less than the confidence intervals, excepted maybe for the system D3. Moreover, linear fits of the results obtained for all the systems are almost the same for the two voices: $y = -0.1309x + 0.9582$ for the "diphone voice" and $y = -0.1285x + 0.9498$ for the "natural voice". One can conclude that the differences between "natural voice" and "diphone voice" are not significant. This indicates that prosodic transplantation is methodologically

relevant as the results look independent of the specific technique used for transplantation.

3.3. Differences between sentences

As only a few (7) sentences are used in this experiment, it is important to check the effect of the sentences. Preference rates by systems and by sentences are plotted in Figure 3.

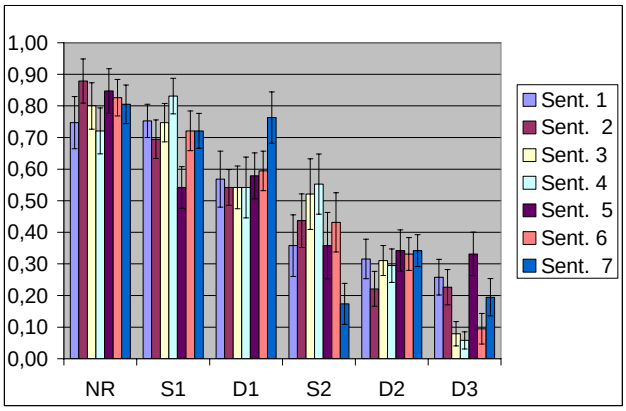

Figure 3: Preference rates by sentences and systems (both voices pooled, error marks are confidence intervals)

Sentences are differing in lengths, syntactic structures and semantic domains. The systems rankings are clearly varying depending on the sentence. For instance, although the global preference order is NR > S1 > D1 > S2 > D2 > D3, note that for sentence 4 the order is S1 > NR > S2 > D1 > D2 > D3 and for Sentence 7 RN > D1 > S1 > D2 > D3 > S2.

Then, for a specific sentence, ranking of the systems can be dramatically modified. However, more than half of the sentences (Sentences 1, 2, 3 and 6) are following the global order for preference rates. More statistical analyses should be performed for a more accurate analysis of the "sentence" factor effect. One can suspect that this factor has a significant effect.

3.4. Differences between subjects

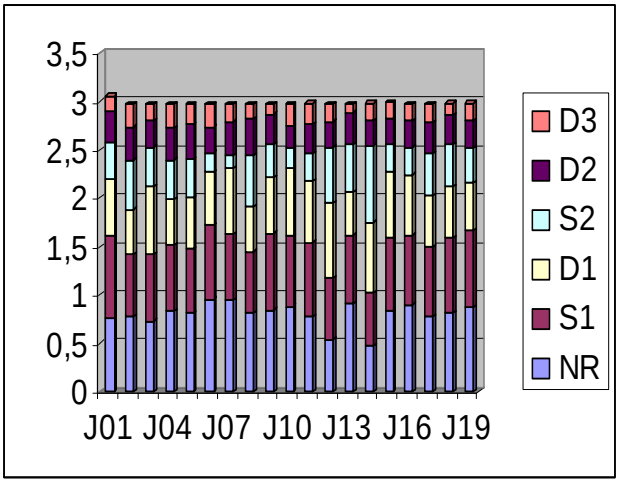

Figure 4: Preference scoring for each judge and for each system (both voices pooled).

A reasonable number of subjects (19) participated in the experiments. On average Global Preference rates clearly indicate the systems ranking. However, a more detailed inspection shows that large differences between subjects are noticeable. Individual preference rates are plotted in Figure 4 (both voices pooled). In this Figure, preference rate of a specific system compared to all the other systems are displayed for each subject (subject 1–19).

As expected, these figures are consistent with the general form of Figures 1 and 2. However, some subjects show outstanding preference rates. For instance subject 14 results give S2 > D1 > S1 > NR. Subject 12 results give D1 > S1 > S2>NR 3. Overal, NR ranks first for sixteen subjects, S1 for one subject, D1 for one subject and S2 for one subject. One can also note the interaction between the effects of the voice and subjects. Between "diphone voice" and "natural voice":

- the ranking of the systems is identical for three subjects (J03, J06 and J07)
- there is one inversion of preference between two systems for eight subjects.
- The three preferred systems are not the same for both voices for six subjects. For five of those subjects, this is because S2 is preferred to D1.

Overall, one can conclude that large intersubject variations exist. In this paper, all the subjects results have been considered, although it could have been possible to withdraw one or two subjects who exhibited atypical results.

4. Discussion

4.1. Methodological issues

First of all, the prosody transplantation test is expensive, both in terms of subjects time (like all subjective tests), but also in terms of data preparation. Data coming from the different synthesis systems must be carefully checked for grapheme-to- phoneme consistency. In French, a same orthographic text may often give different phonemic representations, because of the mute *e*, because of 'liaisons' and so on. Such differences must be avoided, because for prosodic transplantation one must assume exactly the same segmental content.

For preparation of the data, the participant must either adapt his prosodic description to the ".pho" format, which is quite easy, or alternatively he must perform a phonemic alignment (e.g. by using MBROLIGN).

Another difficulty is the average pitch and speaking rates that can differ between systems. Preliminary testing showed that these effects should be neutralized, because e.g. a high pitched voice transplanted on the MBROLA male voice would sound rather unnatural, despite a good prosodic contour. Then all the systems have been pitch scaled to the target voices.

Global Preference rates (i.e. one system against all the other systems) have been used in this paper for analyzing the results. However, this is not the only way to compare the systems. For instance, paired preference rates could have been used as well. A difficulty with global preference rates is that they are percentages that do not sum to one. Computing statistical significance values for these types of data seems not straightforward. Then further analyses are needed in order to strengthen the tendencies found in our results.

In summary, we think that despite its relatively high cost, prosodic transplantation can be recommended for

establishing benchmarks and for comparing the prosody of a specific system to the prosody of other systems and natural voice.
However, this test should be performed along with a diagnosis test in order to pinpoint the weaknesses of the system regarding prosody. It would also be interesting to perform an absolute category rating (ACR) test, using categories relative to prosody. Finally, it would be interesting to compute additional results like the score of one system again another one instead of against all the systems. This would help us in discriminating systems with close results when they are compared to all the systems.

4.2. Conclusions

One of the questions raised by previous experiments using prosodic transplantation was the effect of the diphone base used. This experiment shows no significant differences in the results obtained for the "diphone voice" condition compared to the "natural voice" condition. This is an interesting result because prosodic transplantation can easily be implemented in many languages with the help of publicly available diphone systems like MBROLA.
The global results show the following ranking: NR > S1 > D1 > S2 > D2 > D3. They are consistent with the mean opinion score (MOS) and other Absolute Category Ratings obtained with the same systems in another experiment (Boula de Mareüil *et al*, 2006), for which the same ranking was obtained. These ACR tests are mainly an indication of the global perceived quality. It seems that prosodic quality is highly correlated with the overall perceived quality. Different figures have been obtained for other aspects of evaluation, namely grapheme-to-phoneme transcription (Boula de Mareüil *et al*., 2005) and intelligibility (Boula de Mareüil *et al*., 2006). Prosodic quality seems rather uncorrelated with these important aspects of systems performances. A remarkable feat is the high quality obtained by the two top ranking systems. They are preferred to natural prosody for some sentences. This has already been observed in a previous experiment (Prudon *et al*., 2004): a real speaker can be outperformed by a well-tuned synthesis system, as far as prosodic quality is concerned. A synthetic system can be more lively and seducing than a monotonous natural voice.
Large variations in results are observed among sentences. Although the general tendency is clear, very different results can be found for specific sentences. The perceptual or cognitive bases for prosodic preference are still almost completely unknown, and it is difficult to explain why a specific prosodic pattern would be judged preferable to another. Then, it is strongly advised to use as many sentences as possible in such prosodic transplantation experiments, because large variations between sentences are expected. The effect of sentence length is not clear, but too short sentences should certainly be avoided.
The 19 subjects in our experiments showed a consistent behaviour, except 2 or 3 subjects. Rather large variations in results between subjects are noticeable. Again, it is difficult to infer the mechanisms for prosodic preference. Then, more analysis of the subject behaviours would be needed.
By products of this research are the evaluation method (prosodic transplantation) and the evaluation platform, a valuable resource for subjective tests, which will hopefully be useful for further TTS system evaluation campaigns.

5. Acknowledgements

The EVALDA evaluation campaign is supported by the French Ministry of Research in the context of the Technolangue programme. Thanks to Olivier Deroo for providing the Acapela stimuli.

6. References

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 converters." In *Proc. Eurospeech'05, (Interspeech)*, Lisbon, pp. 1521–1524.

Boula de Mareuil, P., d'Alessandro, C., Raake, A., Bailly, G, Garcia, M.N., Morel, M. (2006), "A Joint intelligibility evaluation of French text-to-speech systems: the EvaSy SUS/ACR campaign" In *Proc of LREC*, Genoa.

Dutoit, T. *et al.* (1996) "The MBROLA project: towards a set of high quality speech synthesizers free of use for non commercial purposes" In *Proc. of ICSLP*, Philadelphia, pp. 1393–1396.

Lamel,, L.F., Gauvain, J.-L., Eskénazi, M. (1991), "BREF, a Large Vocabulary Spoken Corpus for French" in *Proc of Eurospeech*. Genova, pp. 505–508,

Malfrère, F, Dutoit, T (1997) "High Quality Speech Synthesis for Phonetic Speech Segmentation", Proc. Eurospeech '97, pp. 2631-2634.

Prudon R., d'Alessandro C., Boula de Mareuil P. (2004) "Unit selection synthesis of prosody : evaluation using diphone transplantation" In S. Narayanan, A. Alwan (eds) *Text-to-speech synthesis: new paradigms and advances*, Prentice Hall, New Jersey, pp. 203–217.

7. Annex: sentences used in the experiments

1 Les déclarations du président-directeur général de France-loto, monsieur Gérard Colé, mardi 16 janvier sur Europe 1, les ont cependant plongés dans l'embarras.
2 Système d'information opérationnel, partie prenante du processus de production ou système d'information de pilotage et d'aide à la décision.
3 Dans un passé encore récent, certains ont voulu fractionner le CNRS, le transformer en une agence subventionnaire d'une Université alors en crise.
4 En 1987, a commencé la grande aventure du marché continu informatisé.
5 L'Indonésie s'efforce de réunir, le mois prochain à Djakarta, les factions cambodgiennes.
6 Quant aux bénéfices nets, ils atteindront 120 millions dans les comptes 1989.
7 Le gouvernement — la chose est trop méconnue — n'intervient qu'en aval.