Judgement on quality and diagnostic evaluation of synthetic prosody

Serge Santi & Isabelle Guaïtella
Institut de Phonétique d'Aix-en-Provence
Laboratoire Parole et Langage URA CNRS 261

ABSTRACT

Two original methods of evaluation of synthetic prosody have been developped and applied to three prosodic rule generators using four French dialogue oriented applications (plane tickets reservation, remote inquiry of databases, etc.). Our data show a good correspondance between the results of two tests. They are both capable of discriminating between algorithms according to each type of application. Each method seem to meet our expectations, i.e., evaluation of quality only vs diagnostic evaluation. The question of what is to be evaluated must be answered before determining how to evaluate.

INTRODUCTION

Prosody can be considered an independant module of the speech synthesis process. As a consequence, a separate evaluation of both prosodic and segmental rules is needed. For prosody -and even for the segmental level-, despite the interest of the speech technology community (see, for instance, Grice et al., 1991), no standard evaluation method seems to be available yet (Santi, 1992).

The principle according to which the goals of the experiment must determine the choice of the method, seems to be unequivocal. For instance, if different systems (or algorithms) are to be compared, a quality test based on satisfaction scores (see, for instance, Pavlovic et al., 1990) or a pair comparison procedure have to be conducted. On the other hand, if evaluation is carried out in order to provide information about eventual defaults of the rules, then a diagnostic evaluation is needed.

Two original methods of evaluation of synthetic French prosody (Test 1: Localisation of prosodic defaults by underlying misfunctionning sequences, Test 2: Evaluation of quality based on the satisfaction criterion) have been developped and applied to three prosodic rule generators using four dialogue oriented applications. These methods and their main results will be briefly presented here.

METHOD

Speech material

Four representative dialogues corresponding to four applications were considered: "Route" (Road): delivery of messages about road traffic (no human speaking intervention); "Camif": shopping orders by telephone (no human speaking intervention); "SNCF" (Train): train ticket reservation (with human speaking intervention); "Avion" (Plane): plane ticket reservation (with human speaking intervention).

Three prosodic rule generators (also called prosodic "styles") were tested: Multivoc, Cnetvox-lecture, Cnetvox-dialogue. The segmental continuum (male voice) was synthesised by a TTS rules system develloped by the French Telecom (CNET at Lannion, France). Recordings of the user -for the two last applications- were recorded in Aix-en-Provence by a female speaker in order to avoid possible ambiguity between the human speaker and the machine.

All dialogues were segmented into pragmatic units called "blocks". All blocks have semantic, syntactic and pragmatic (interactional) coherence. As a consequence the size of the block is highly variable (from 8 words to 84 words, from a single sentence to 8 speech turns).

Example:

<u>block Avion 3</u>

Machine:	quel jour desirez-vous partir
Human:	*le cinq avril en fin d'après-midi*
Machine:	le cinq avril vers quelle heure
Human:	*à partir de dix-sept heures*

The method used in test 1required a written version of the dialogues. Because of the close relationship existing between punctuation and prosody (Guaïtella & Santi, 1992), no punctuation marks were used.

Subjects and hardware

A first group of thirty listeners took part in test 1 and fifteen others in test 2. They were all native French students (male and female in equal proportion) and were paid for the test.

Speech material was presented to listeners by means of headphones (AKG, K240). Storage and restitution of audio stimuli was completed with an Intel 386/25 PC micro-computer. The human voice was recorded on a SONY Digital Audio Tape. Tests 1 and 2 were performed individually in a sound-proof chamber.

Test 1: Diagnostic procedure of localisation of misfunctionning sequences

Subjects had to listen carefully to each block and to concentrate on the prosody. In a second hearing they were asked to to underline all eventual sequences they judged to be not or hardly acceptable. Written and oral explanations and a pilot test were proposed to the subjects before the test. All subjects listened to all blocks in a specific random order but a single prosodic version was presented to each subject. The duration of test 1 was about 25 minutes. This method is similar to that used by Hirst et al. (1991)

Test 2: Evaluation of quality based on satisfaction scores

This method is based on methodology proposed by the SAM (Speech Assessment Methodology) working group on prosody evaluation (Grice & Hirst, 1991). We also used the SOAP software to pilot the test (Howard-Jones et al., 1991).

Listeners had also to listen carefully to each block and to concentrate on prosody. After each block they were asked to give a satisfaction score based on prosody only (they were asked not to judge segmentals). A scaling method was prefered to magnitude estimation (see Pavlovic et al., 1990). All other elements were similar to test 1 exept the fact that subject listened to the three prosodic versions according to the Latin Square method.The duration of test 2 was about 25 minutes.

RESULTS

Test 1

Due to the nature of the data, a manual exploitation of the answers was carried out. For each block two scoring methods were used: 1- Number of underlinings, 2- length of underlining distance, in mm (see Hirst et al., 1991). Because of the non-homogenous sizes of the blocks relative values were calculated (percentage of underlined text -distance-, or percentage of underlined words).

Two analyses of variance were carried out:

ANOVA 1- factors: prosodic style and application, response: number of underlinings.

ANOVA 2- factors: prosodic style and application, response: percentage of underlined text.

For ANOVA 1 both factors are highly significant (prosodic style: $p < .0003$; application: $p < .0001$) but the interaction is not ($p <.1222$). For ANOVA 2 both factors and interaction are significant (prosodic style: $p < .0001$; application: $p < .0001$; interaction: $p <.0232$). The incidence table of ANOVA 2 shows that some prosodic styles seem to be better adapted to some applications (for instance: Route and Cnetvox-lecture) (table 1)

Table 1. *Incidence table of ANOVA 2: interaction between "prosody" and "application" factors (percentage of underlying text is taken into account).Up: population, low: percentage.*

	application:	avion	route	sncf	camif	Totals :
prosody	Multivoc	100 21,44	90 11,53	100 19,24	120 9,2	410 15,16
	L-cnetvox	100 13,34	90 7,9	100 12,4	120 9,13	410 10,68
	D-cnetvox	100 13,33	90 12,5	100 12,46	120 7,65	410 11,27
Totals :		300 16,04	270 10,64	300 14,7	360 8,66	1230 12,37

A significant but weak correlation between the two types of scores (number of underlinings and percentage of underlined text) does not allow us to consider only one of these scores. We also noticed that individual variability was important. Listeners used different individual strategies in underlining. For more details concerning the results of test 1 see Santi & Guaïtella (1992a).

Test 2

Here too two analyses of variance were computed.
ANOVA 3- factors: prosodic style and application, response: scores.
ANOVA 4- factors: prosodic style and application type, response: scores.

The "application type" factor is a combination of the applications. The applications Route and Camif where no human intervention by voice is involved were labelled as *monologue*. The two others application (Train and Plane) are grouped into the *dialogue* category.

For ANOVA 3 both factors and interaction are significant (prosodic style: p < .0001; application: p < .0229; interaction: p <.0164) (Table 2). ANOVA 4 also shows a signicant effect of both "prosody" and "application type" factors but the interaction is not significant (p <.4969).

Table 2. *Incidence table of ANOVA 3: interaction between "prosody" and "application" factors (scores).Up: population, low: percentage.*

	application:	route	camif	sncf	avion	Totals:
prosody	Multivoc	45 9,733	60 8,567	50 8,64	50 10,46	205 9,302
	L-Cnetvox	45 12,733	60 11,95	50 12,96	50 12,4	205 12,478
	D-Cnetvox	45 11,356	60 12,6	50 12,8	50 13,56	205 12,61
Totals:		135 11,274	180 11,039	150 11,467	150 12,14	615 11,463

The results of test 2 clearly show that Cnetvox-lecture and Cnetvox-dialogue are judged better by listeners than Multivoc. Hierachy between Cnetvox-lecture and Cnetvox-dialogue is harder to establish but we can notice that Cnetvox-dialogue is prefered for dialogue applications and Cnetvox-lecture best evaluated for monolog applications. This result can be explained by the fact that Cnetvox-dialogue takes into account some dialogic aspects of intonation. On the contrary Cnetvox-lecture is the standard prosodic output of the TTS system and is based on a "reading" model of intonation.

DISCUSSION

Results of test 1 are consistent with the results of test 2. While the two methods are quite different, the results often lead to the same conclusions:

- performance of Cnetvox-lecture and Cnetvox-dialogue are always superior to Multivoc.
- these prosodic styles are hardly discriminable and are dependent on the application in which they are used.
- the coherence between the results of test 1 and test 2 validate both methodologies...

In fact, the diagnostic capabilities of test 1 have been exploited but are not described here (see Santi &Guaïtella, 1992b). Test 1 is far more difficult to carry out than test 2 but is not suited to the same goals. If only subjective evaluation in order to discriminate among different algorithms is needed test 2 is sufficient and more efficient. However, if diagnostic information about misfunctionnings of the algorithms are of interest, then test 1 can be quite useful. No evaluation method can tell directly what to do but a well chosen methodology may be capable of telling you on what aspects of the system have to be improved. Concerning methodology itself we claim that the best method should be that closest to the real situation of communication. Even if in a test situation the listener cannot be considered as a user but rather as an observer, the coherence of what he observes is fundamental. As a consequence we maintain that isolated sentences or repetitive speech turns (i.e without any pragmatic context nor semantic coherence) do not constitute a good material for evaluation tests. Larger units such as the blocks used in our tests are certainly much more suited to keep the listener's attention on *what* is said even if he has to concentrate on *how* it is said.

AKNOWLEDGEMENTS

This research was financed by France Télécom (CNET, Lannion; Convention CNET/IPA CB 500 437). Thanks to D. J. Hirst, R. Espesser and C. Pavlovic.

REFERENCES

Grice M., Hirst D.J., 1991, "The evaluation of prosody in text-to-speech systems in a number of languages", *SAM Internal Repport,* Esprit Project 2589, So.2 (part 1).

Grice M., Vagges K., Hirst D.J., 1991, "Assessment of intonation in text-to-speech synthesis systems - A pilot test in english and italian", *Proceedings of the Eurospeech Conference,* Genova, vol.2, 879-882.

Guaïtella I, Santi S., 1992, "The punctuation and perception of read and spontaneous prosody: an application to speech synthesis", in: G. Bailly, C. Benoit & T. Sawallis (eds), *Talking machines: Theories, models and designs*, Elsevier, North Holland, 351-366.

Hirst D.J., Nicolas P., Espesser R., 1991, "Coding the f0 of a continuous text in French: an experimental approach", *Proc. of the XIIth I.C.Ph.S.*, Aix-en-Provence, 234-237.

Howard-Jones P.A. & the SAM Partnership, 1991, "'SOAP' - A speech output assessment package for controlled multilingual evaluation of synthetic speech", *Proceedings of the Eurospeech Conference,* Genova, vol.1, 281-283.

Pavlovic C.V., Rossi M., Espesser R., 1990, "Use of magnitude estimation technique for assessing the performance of text-to-speech synthesis systems", *J. Acoust. Soc. Am.*, 87, 373-382.

Santi S., 1992, "Methodes d'evaluation subjective de la composante prosodique en synthèse vocale", *Prépublication des actes du Séminaire Prosodie,* Aix-en-Provence, 36-46.

Santi S., Guaïtella I., 1992a, "Evaluation de la qualité de la prosodie de synthèse en situations de dialogue", *Document CNET/IPA 92-06.*

Santi S., Guaïtella I., 1992b, "Localisation et analyse de défauts prosodiques en synthèse du dialogue", *Document CNET/IPA 92-11.*