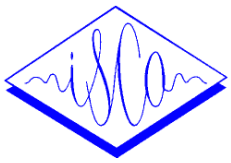

Fourth ISCA ITRW on
Speech Synthesis (SSW-4)
Perthshire, Scotland
August 29 - September 1, 2001

Prosody evaluation as a diagnostic process: subjective vs. objective measurements

Albert Rilliard & Véronique Aubergé

Institut de la Communication Parlée
Université Stendhal, Grenoble, France
{rilliard, auberge}@icp.inpg.fr

Abstract

A set of perception experiments, using reiterant/lexicalised speech, were designed to carried out a diagnostic of the prosodic function of segmentation/hierarchization. Both natural and synthetic intonation were evaluated. Then, several dissimilarity measures – correlation, root-mean-square distance and mutual information on the acoustic parameters (Fo, syllabic duration and intensity) – were applied to match the perceptive results. This objective vs. subjective comparison underlines which acoustic keys are used by listeners to judge the adequacy of prosody in performing a given function such as demarcation.

1. Introduction

Speech synthesiser evaluation is a recurrent theme in the field of speech technologies. However, most of the evaluation procedures are based on global criteria of the synthetic voice quality, such as the naturalness, the pleasantness, the ease of listening... If such criteria adequately reflect the appreciated qualities on the client's view, they do not allow, on the designer's view, a diagnostic measure of the structures implicitly or explicitly implemented in the synthesiser; and neither can they directly report the adequacy of a synthetic voice to a given situation.

As speech synthesisers evolve, we can recall Campbell's claim: "...whereas early speech synthesisers were designed primarily to be 'reading machines', with the emphasis on intelligibility, the coming generation should be thought as 'speaking machines' instead, with more emphasis on naturalness of speech production." (Campbell, 1998, p. 17), and then consider speech style as a basic information, that organises the underlying structures. Evaluation in this context has to report on:

- Criteria that characterize the adequacy of a style to a situation; such criteria will probably be analogous to those used for natural voices, as the clients' requirements are the same (e.g. lexical and syntactic intelligibility for weather forecast reports; comfort, peacefulness for airport announcements);
- The synthesizers' failure in generating speech that respects a given style criterion; or more precisely, a diagnosis of the incompetence in generating adapted structures for a given style.

In this view, a long-term aim for prosody evaluation is to inventory precisely all the different functions performed by prosody, in order to design measurement procedures to evaluate the realisation of each function by a given set of prosodic parameters. It will then be possible to map, for one style, an "ideal" realisation of each function in each situation, with a corresponding natural prosody, used as a reference that could be used to rate the competence of one or more speech synthesisers.

As most speech synthesiser are built as a succession of dedicated modules (van Santen, 1997), the use of such a prosodic reference seems easy to apply. With such a modular construction of synthesisers, many approaches of evaluation are intended to produce, for designers, a diagnosis of the errors made by each module (cf. the results of the evaluation experiments presented in Yvon et al., 1998). But the most of the works published on evaluation deal with segmental intelligibility, leaving the field of prosody evaluation nearly empty (Fourcin, 1992).

However, a modular approach of evaluation should not be justified only by the design of synthesisers. It should be possible to "isolate" each linguistic structure implied in the human perception system, i.e. to be able to evaluate natural speech (the reference) in the same way as synthetic speech is – being aware that freezing all the other linguistic modules to observe the last one may induce a bias, even more important in synthetic speech than in natural speech (cf. discussion in Yvon et al., 1998).

The evaluation of linguistic structure's performances can be conceived at several levels. One can (1) evaluate the message ability to perform adequately the structures that carry a given function; or (2) measure, in both natural and synthetic speech, the (non)achievement of a function as a stylistic rating, and an assessment of the robustness of this function in regard to the situation's constraints. This work is linked with the first of these two points: trying to measure the relative contribution of prosody to a given linguistic function realisation, the segmentation and hierarchization of utterances.

This study stands at the end of a methodological chain, that aims at prosody modelling (Aubergé, 2000): A hypothetico-deductive approach, beginning with the formulation of theoretical hypothesis extracted from a model, followed by an experimental validation of this model. Such an approach has already been used by Martin (1980), to test the congruence hypothesis between syntax and prosody, and is similar to those used at the ICP (cf. Aubergé & Bailly, 1995). Starting from corpora (based on strong theoretical hypothesis), the chain is continued by a first order (Aubergé, 1991) or second order (Morlec, 1997) statistical analysis emerging onto a generation model. This generation model is supposed to produce at least the same utterances as those of the learning corpus, and expected to be able to generalise its competencies to others structures (Morlec, 1997). The evaluation proposed here consists in a validation of the capacities of this model, as compared to the performances of the natural prosody (the reference) contained in the corpus.

associate a natural (well-formed) reiterant prosody with a synthetic lexicalised stimulus. Such a condition is necessary to validate the paradigm consistency: if a reiterant natural prosody (adequate for a given function) is associated to a lexicalised stimulus (judged adequate for the same function in C5 condition), it can be assumed that C6's results must be coherent with C5, and both stimuli therefore associated by listeners. This check is interesting because the reiterant speech paradigm is not clearly ecological and because the experiment imposes an explicit use of metaknowledge information to listeners.

2.1.1. Experimental Protocol

For each condition, the text of the utterance is given to the subjects. Then they hear a reiterant stimulus and, depending on the condition, a lexicalised acoustic stimulus half a second after (for conditions C2, C4, C5, C6). Subjects answer by "Yes" or "No" depending on whether the reiterant stimulus might be convenient or not for the sentence displayed on the screen. They then give a confidence score to their answer, and validate the answer by clicking on an "OK" button, and start with another pair. The reaction time between the end of the acoustic stimulus presentation and the "OK" button click is recorded.

A $S_{13}<C_6>*P_X$ experimental design was used to analyse by a two-ways ANOVA the experimental data. S stands for the subjects of the experiments (13 for each condition); C stands for the six different experimental conditions; and P stands for the pairs of stimuli (representing the syntactic variation). As stimuli are grouped together with respect to their syllabic length, and the individual syntactic structure of each sentence differs from one sentence length to another one, results are analysed separately for each stimuli length. The number of pairs built for the experiment depends on the number of same length sentences. The longer a sentence, the more complex the syntactic variation: it explains why longer sentences than shorter ones have been selected.

- The two 5-syllable sentences produced 4 pairs of stimuli;
- The three sentences for each 6-7-8 and 9-syllable sentences produced 9 pairs of stimuli per length;
- The four 10 and 11-syllable sentences produced 16 pairs of stimuli each.

Referring to the experimental design, since the analysis must be held separately for each stimuli length, the results will be described according to the C (condition), and P (syntactic structure of the pair of stimuli) variables – as the Subjects are supposed not to affect the results.

2.1.2. Stimuli construction

The natural lexicalised stimuli were read aloud by a trained speaker; the synthetic lexicalised stimuli were produced by transplanting the synthetic prosody (from the ICP TTS system) on a recto-tono read (lexicalised) sentence, using a TDPSOLA algorithm; the same procedure was applied for reiterant sentences: both natural and synthetic prosody were transplanted on a read "mamama" sentence (produced by the same speaker as for the synthetic lexicalised version).

Using such a procedure should avoid some artefacts which could be induced by the concatenation of the same diphone when using completely the TTS system for generating the synthetic reiterant speech. "Natural" reiterant stimuli were produced using a transplantation method, and not directly a human speaker, in order to avoid a possible bias of the experiment, due to different qualities of speech signal between really natural reiterant signal and a synthesised reiterant signal. Before performing this experiment, we carried out a preliminary exploration test (Rilliard & Aubergé, 1998), based on the same procedure as conditions C1 and C2, but using completely natural reiterant and lexicalised stimuli. The stimuli of this first test were produced by two speakers (a female and a male). The results of this preliminary test are used to chek a possible bias of the transplantation method.

Indeed, such a transplantation procedure is not flawless: for some natural lexicalised stimuli, a silent pause is produced by the speaker, that is not reproduced in the reiterant stimuli. These stimuli have a very specific rhythmic structure possibly introducing a bias in the experiment.

2.2. Results

Results are both extracted from an ANOVA study, used to analyze the divergences between the different conditions or different stimuli in the same condition, combined with post-hoc tests to refine the analysis of significant effect; and from a chi square test, used to determine if the association answers differ from a 50% chance distribution.

The global results are presented hereafter in the figure 2 for the global association scores and the chi square test. A complete analysis of this results can be found in Rilliard (2000).

For the analysis, the pairs of stimuli presented to listener are group together in six categories, in accordance with the opposition of syntactic structure presented between each stimuli pair:

- same sentences (homogeneous pairs);
- same location of the major syntactic boundary - groups' level and nature identical
- same location of the major syntactic boundary - same level / different nature
- same location of the major syntactic boundary - different level
- shifted syntactic boundary - same level / different nature
- shifted syntactic boundary - different level

Each of these six categories group pairs of sentences which are supposed to propose comparable syntactic differences, each categories are ranked from the "closest" (in term of syntactic divergences) to the more distant.

2.2.1. Acoustic analysis of prosodic parameters

The Fo, syllabic and GIPC (Marcus, 1976) duration, and intensity parameters are extracted from the stimuli, natural or synthetic, reiterant or lexicalised. The acoustic parameters of reiterant sentences are then compared to those of lexicalised sentences, reusing the stimuli pair of the experimental condition C2, C4, C5 and C6. The comparison is made by means of root-mean-square distance and the correlation between each kind of parameters. Next, the acoustic distance found between reiterant and lexicalised stimuli are compared (correlation measure) to the association score of the corresponding pairs.

For the C1 and C3 condition, association scores are compared to the acoustic distance found respectively with the stimuli couple of condition C2 and C4.

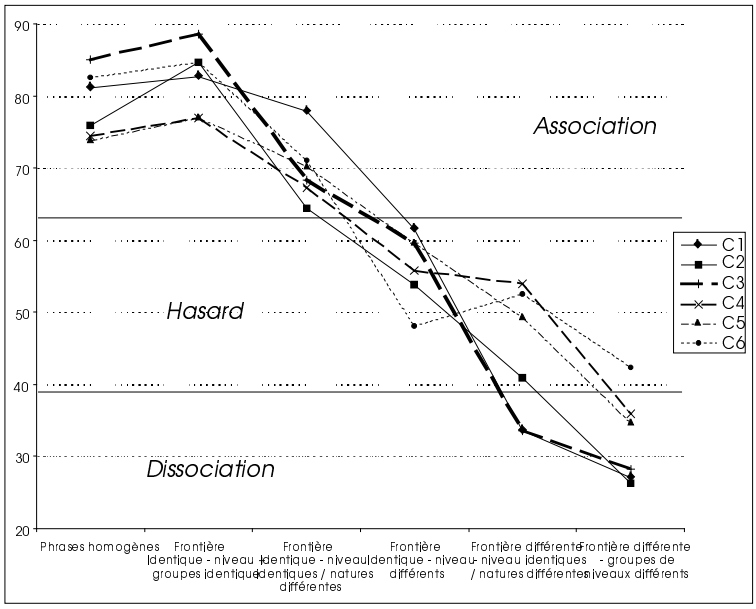

Figure 2: *mean values of the association scores, for each six categories of stimuli pairs. The two horizontal bars point out the boundary beyond which the association scores are significantly divergent from chance*

2.2.2. Analysis

The result of this experiment enhance the global coherence found between the different experimental condition, the result of which are generally similar, validating the good global performance of synthetic prosody, for the demarcation function.

The factor that explain the major part of the variance is the variation of syntactic divergences amongst the stimuli pairs. This result is very important, as it show the listeners ability to catch information consistent with the demarcation function. Moreover, it can be checked that the association scores are coherent with the classical description of the demarcation function for French prosody (cf. Hirst & Di Cristo (Eds.), 1998; and Campbell, 1993a and 1993b). The association result summarised the relative importance of the syntactic divergences represented in each of the six groups of stimuli pairs, as shown by the figure 2.

Then, a detailed comparison of synthetic and natural prosody performances emerge to a precise diagnosis of the synthesiser's strengths and weaknesses:

- basic syntactic structures (Subject Verb, or Subject Verb Object) are very well performed by the generator, reaching the over-learning for this structures;
- rare syntactic structure (in the learning corpus) are mis-produced by the synthesiser (structure based on enumeration and long nominal group);

From the acoustic analysis, information retrieved concern:

- the Fo, always highly correlated with perception results, either for natural and synthetic one – result that bears out the primary role of this parameter in the demarcation function, and the good efficiency of the synthesiser to model the Fo;
- the duration values also feet with association results, with higher scores for GIPC than for duration measure – but the correlation are not as good as they are in the preliminary experiment, that could be a bias of the transplantation procedure, and the synthetic duration are not as efficient as natural ones.

3. Conclusion

As already seen at the beginning, the diagnostic evaluation of prosody, if still almost unexplored, is already rich of new paradigms proposals, or reuse of already known paradigms.

The experience gained during the AUPELF-UREF action for the evaluation of synthetic prosody (which consist, at the moment, in a list of possible paradigms), confirm this impression: already existing protocols that could be directly apply to the prosody evaluation are multiples (test based on MOS scales, Sentences Verification Task, SUS, disambiguation by mean of prosodic information). The experiment described here complete this list, with a special emphasis on a task dedicated to the localisation of miss-produced synthetic structures, in order to propose a diagnosis evaluation to the system's designers.

Thus, this experiment try to directly observe the chosen linguistic function (here the demarcation function), through a metalinguistic association task, already used and validated (see e.g. Liberman & Streeter, 1978, Larkey, 1983), but complex to realise. It is mainly here a feasibility study for such an experimental protocol. The prosodic structure are tested here without any constraints: the pure prosody is directly compared to the syntactic structure1, and not to a referential prosody, that could introduce a "norm" of the maximum possible performances for prosody, as it is the case in the experiment held by Morlec et al. (1998; see Rilliard, 2000, for a detailed comparison). In the paradigm proposed here, the subjects' answers can be better for

1 At least for the experimental conditions C1 and C3, where the only acoustic stimuli is the reiterated one.

synthetic stimuli than for natural ones, for some specific syntactic structures. Such a result can be interpreted as a the production of a stereotypical structure, over-learned by the generation model, to the detriment of others structures, the performance of which is under-rated by listeners, in comparison to natural performances. The analysis of this experiment underline the relative performances of natural and synthetic prosodies, similar for some structures, and divergent for some others, leading the experimenter to a precise diagnosis of the generation system.

As explained in introduction, This compared evaluation is an important part of the ICP process of prosody modelling. The version of the synthesiser tested here was established specifically to produce prosodic attitudes (Morlec, 1997), but not the demarcation function, only quickly learned, on a small corpus. Since, a new version was realised by G. Bailly, based on a more complete learning corpus, specifically dedicated to the segmentation hierarchisation function (Aubergé, 1991). Concerning the diagnosis of the first version of the synthesiser (Morlec, 1997), the following point can be underlined, concerning the syntactic structures, and the quality of their reiteration:

- excellent performances for Subject Verb Object structures, with very strong demarcation capabilities for stimuli of length from 5 to 9 syllables – performances increased to the over-learning of such structures;
- gaps concerning syntactic structures not represented in the learning corpus, as enumeration or long nominal groups;
- good modelling of the fundamental frequency;
- some problems with the generation of segmental duration: too isochronous?
- no pertinent results concerning the intensity parameter.

With regards to the information that can be extracted from such an experiment on the cognitive treatment of prosody, the capacity of listeners to extract the demarcation cues from the incomplete input flow, containing only pure prosody information, can be remembered. This result comforts the hypotheses of a modular treatment dedicated to prosody, or at least does not disprove this hypothesis. By comparing this results with the hypotheses expressed by Gerard & Dolgër (1996), about auditory span dedicated to prosodic information storage, mechanism specialising the so called “articulatory loop” (Baddeley, 1986), a temporal boundary can be fix to the treatment of prosodic information, corresponding to approximately 11 or 12 syllables. And this even if our experimental corpus is biased, the long sentences carrying systematically complex syntactic structure. Indeed, the experiment performed by Rolland (2000), where the syntactic complexity is kept the same for all stimuli length, shows the same limit in the listeners’ ability to use information from delexicalised stimuli.

Moreover, the nature and the amount of syntactic information that the listeners “retrieved” by means of prosody, information not systematically correlated in value and in place with the syntax, confirms, in our view, the non-hierarchical functioning of the prosodic module, in comparison to syntax. This is the sole functioning that can integrate a restraint of coherence between syntax and prosody for the same demarcative function, and a degree of freedom in the realisation (functioning described by means of “structural rendezvous” by Aubergé, 1991), and then allowed to answer both to a morphologic criterion (robustness) and to a stylistic criterion (organisation of structures), already noted by Campbell (1993a).

4. Acknowledgements

We are deeply grateful to Gérard Bailly and Yann Morlec for their fruitful advice on the theoretical problems raised by this experiment.

5. References

[1] Aubergé V., 1991. “la synthèse de la parole : des règles au lexique”. PhD Thesis, Grenoble, France.

[2] Aubergé V., 2000. “Modélisation de la prosodie par formes globales : amont ou aval de la phonologie tonale ? L'exemple d'un modèle développé à l'ICP” Actes des 23e Journées d’Études sur la Parole, Aussois, France, 281-284.

[3] Aubergé V. & Bailly G., 1995. “Generation of intonation: a global approach” Proceedings of EuroSpeech’95, Madrid, Spain, 3, 2065-2068.

[4] Baddeley A.D., 1986. “Working memory” Oxford University Press.

[5] Campbell N., 1993a. “Automatic detection of prosodic boundaries in speech” Speech Communication, 13, 343-354.

[6] Campbell N., 1993b. “Durational cues to prominence and grouping” ESCA workshop on prosody - Lund university working papers, 41, 38-41, Lund, Sweden.

[7] Campbell N., 1998. “Where is the information in speech?” Proceedings of the 3rd ESCA/COCOSDA International Workshop on Speech Synthesis, Jenolan Caves, Australia, 17-20.

[8] Fourcin A., 1992. “Assessment of synthetic speech” In Bailly G., Benoît C., and Sawallis T.R. (Eds.), *Talking Machines – Theories, Models and Designs*, Elsevier Science, Amsterdam, 431- 434.

[9] Gérard C. & Dolgër N., 1996. “Taille des fenêtres perceptives, empan de la mémoire auditive”. XXIème Journées d’Étude de la Parole, Avignon, France, 59-62.

[10] Hirst D. & Di Cristo A., (Eds.), *Intonation systems: a survey of twenty languages*, Cambridge University Press.

[11] Larkey, L.S., 1983. “Reiterant speech: an acoustic and perceptual validation”. Journal of the Acoustical Society of America. 73(4), 1337-1345.

[12] Liberman, M.Y. & Streeter, L.A., 1978. “Use of nonsense-syllable mimicry in the study of prosodic phenomena”. Journal of the Acoustical Society of America. 63 (1), 231-233.

[13] Martin P., 1980. “De la non congruence entre les structures syntaxiques et prosodiques” Travaux de l’Institut de Phonétique d’Aix, 7, 319-339

[14] Marcus S.M., 1976. “Perceptual Centers”. Ph.D. Thesis, Cambridge University, UK.

[15] Morlec Y., 1997. Génération multiparamétrique de la prosodie du français par apprentissage automatique. PhD Thesis, Institut National Polytechnique de Grenoble, France.

[16] Morlec, Y., Rilliard, A., Bailly, G. & Aubergé, V., 1998. “Evaluating the adequacy of synthetic prosody in signalling syntactic boundaries: methodology and first results”. Proceedings of the first International Conference on Language Resources and Evaluation. Granada, Spain, 647-650.

[17] Rilliard A., 2000. " Vers une mesure de l'intelligibilité linguistique de la prosodie – évaluation diagnostique des prosodies synthétique et naturelle" PhD Thesis, Institut National Polytechnique de Grenoble, France.

[18] Rilliard A., Aubergé V., Bailly G. & Morlec Y., 1997. "Vers une Mesure de l'Information Linguistique Véhiculée par la Prosodie", Proceeding of FRANCIL'97, Avignon, France, pp. 481-487

[19] Rilliard A. & Aubergé V., 1998. "Reiterant Speech for the Evaluation of Natural vs. Synthetic Prosody". Proceedings of the International Congress of Spoken Language Processing, Sydney, Australia, pp. 675-678.

[20] Rolland G., 2000. "La pertinence psycho-acoustique du syntagme accentuel en français" Mémoire de DEA Signal, Image, Parole, Télécoms, Institut National Polytechnique de Grenoble, France.

[21] Van Santen J.P.H., 1997. "Prosodic modeling in Text-to-Speech synthesis" Proceedings of EuroSpeech'97, Rhodos, Greece, KN 19-28.

[22] Yvon F., Boula de Mareüil P., d'Alessandro C., Aubergé V., Bagein M., Bailly G., Béchet F., Foukia S., Goldman J.F., Keller E., Oshaughnessy D., Pagel V., Sannier F., Véronis J. & Zellner B., 1998. "Objective evaluation of grapheme to phoneme conversion for text-to-speech synthesis in French" Computer Speech and Language, 12 (4), 393-410.