Towards a Linguistic Validation of a Prosodic Generation Model

Albert Rilliard & Véronique Aubergé

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

Abstract

This paper deals with an approach of prosody evaluation and modelisation, tested together a prosodic model and a prosodic generator, to produce a compared diagnostic of their linguistic competencies. Three evaluation experiments are succinctly described. The first one consists in an online measurement by listeners of the prosodic quality; the second one uses tuned stimuli to test the differences perceived by subjects, and the last one uses delexicalised stimuli to present separately a pure prosodic stimulus and a syntactic stimulus. Their efficiency and disadvantages are analysed along a compared analysis of their results.

1. Introduction

On the one hand, numerous prosodic models have been proposed to describe a number of languages [8]. To compare the theoretical proposals to the real performances, some models are implemented in order to generate simulated speech stimuli (e.g. [3], [11]). On the other hand, text-to-speech systems are evolving from "reading machines" to "speaking machines" (according to [5]), with increasing segmental quality and mainly suprasegmental quality. For French in particular, the synthetic prosody of TTS systems is explicitly based on prosodic models (e.g. SyntAix, the LPL TTS system based on Hirst & Di Cristo model; the KALI system based on Lacheret model [18]; Genève TTS système based on Mertens Model [12], ICP TTS system based on Aubergé model [2]...). It has to be noted that stochastic approaches are very few in TTS, even if models are "data-learned" (more than "data-driven" – see already [19])

In this view, evaluating synthetic prosody in order to diagnose and consequently to improve it, is quite the same problem as validating models' performances: evaluation can be considered to be linked to the model relevant for the TTS system, and it can be related to the compared measurement of natural and synthetic (simulated) performances to realize an explicit linguistic function.

Evaluation in this context has to report on (i) the message's ability to perform adequately the structures that carry a given function; or (ii) the measure, in both natural and synthetic prosody, of the (non)achievement of a function as a stylistic rating, and an assessment of the robustness of this function in comparison with 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.

It is in this scheme that this paper is embedded, along with the prosodic modelisation chain used at the ICP [2]: A hypothetico-deductive approach, beginning with the formulation of theoretical hypotheses extracted from a model, followed by an experimental validation of this model. Starting from corpora (based on strong theoretical hypotheses), the chain is continued by a first order [1] or second order [13] 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 other structures [13]. The proposed evaluation consists in a validation of the competencies of this model, as compared to the performances of the natural prosody (the reference) contained in the corpus. In this view, it is interesting to link a prosodic shape with a given function, with a rated efficiency, for two reasons: (i) to measure the relative importance of prosody in the perception of the studied linguistic function; and (ii) to validate the appropriateness of a prosody produced by a synthesiser to the same given linguistic function, and then to validate the prosodic model underlying the synthesiser.

In this scheme, three experiments were performed (described in [15]), using different experimental paradigms, to test the perceptive adequacy of prosody (either natural and synthetic) to a linguistic function: the segmentation and hierarchization one.

Hereafter those experiments are compared, trying to list their strength and weaknesses in achieving this goals.

2. Perception experiments

2.1. Situation of the problem

The basic problem is to get from listeners linguistic information on the adequacy of the model's parameters to the demarcation function (that is to segment the utterance and to hierarchize the segments), through perception experiment. Three experiments, exemplifying three different possibilities of dealing with such a problem will be succinctly described and then discussed hereafter. Their concepts consist in:

- For the first one, leaving all the responsibilities to listeners, asking them to determine on line if the prosody is well-formed for the sentence.
- For the second, leaving no responsibilities to listeners, modifying the proposed stimuli to present them exactly the same task on different (and possibly incoherent) prosody.
- For the third, using a middle way that consists in proposing to listeners separately prosody and syntax, and to ask them to associate both if they feel it is correct.

After the perception experiment, an acoustic analysis is systematically performed on the stimuli, and then these objective results compared to the subjective ones.

2.2. Online subjective validation

This experiment is based on a work done by [7], who asked listeners to underline unsatisfactory portions of a text, while they were listening to them. They found that the actual length of underlined text was highly correlated to the listeners' global evaluation of the naturalness of the passages.

In the present experiment, the complete text of 20 passages of 5 semantically linked sentences are used. This multi-speaker corpus is extracted from the EUROM1 database. Each passage is displayed on a computer screen, and is read aloud by two synthesizers, while the pronounced word is dynamically selected during its pronunciation, almost like in a "karaoke" session.

Listeners hear the passages, and have to click onto the words when the prosody is judged inadequate (resulting in a local evaluation of prosody). At the end of each passage, they give an overall quality rating for each passage(resulting in a global evaluation of each passage). A more detailed description and analysis can be found in [9]; and [15]. All actions made by subjects are recorded by the computer.

An acoustic analysis of the synthetic stimuli is performed, and then compared to the analysis of the natural speech corpora. The acoustic parameters extracted from the analysis are the fundamental frequency, the syllable duration and intensity. Two acoustic distances (the root-mean-square distance, and the correlation) between synthetic and natural stimuli are calculated.

Those objective distances are compared to the subjective distance obtained from the perception analysis, using the correlation coefficient between the two entities.

2.3. Stimuli with incoherent prosody

The second experiment plans to test the sensitivity of both natural and synthetic prosodies to perturbations in their demarcation function (for a complete description, see [14] and [15]). Also, the corpus is based on a set of sentences with similar phonotactic and phonetic dimensions, but fulfilling different demarcation functions. Stimuli are organized by pairs of sentences, listing a set of minimal pairs of syntactic oppositions.

For each sentence of the corpus (in its natural or synthetic version), the acoustic parameters of prosody (Fo, duration, intensity) are extracted. Then, stimuli are constructed by transplanting the set of acoustic parameters of all sentences in the corpus onto each sentence the same length. It results in either:

- a coherent stimulus (a sentence with its original prosody)
- an incoherent stimulus (a sentence with the prosody from another sentence)

All the possible pairs of a coherent plus an incoherent sentence are made, in order to test the perceptive degradation induced in the incoherent stimuli.

Listeners are asked to judge which item is the most adequate for a neutral reading of the sentence written on the screen. If they are not able to choose the best sentence, they are allowed to answer that both or none of the sentences are adequate.

2.4. Reiterant speech

The third experiment aims at measuring the "linguistic intelligibility" of prosody for the demarcation function. Stimuli are here reduced to prosody, without any other possible linguistic access (using speech reiterated with the canonical syllable /ma/, see [10]). A preliminary experiment with reiterant stimuli produced by humans was performed [16], in order to test the ability of listener to deal with such stimuli.

For this experiment, 22 sentences were produced by analyzing and resynthesizing natural and synthetic prosodies on recto-tono "mamama" sentences. They reflect the opposition of major vs. minor boundaries at the same location, and the occurrence place variation of two equivalent boundaries along the utterances (see [15] for a complete description).

Listeners are asked to associate a reiterant stimulus to a syntactic reference. Depending on the experimental condition, the first part of the pair of stimuli (that is the reiterant stimulus) is either natural or synthetic; the second part of the pair (that is the syntactic reference) is presented through a text displayed on a screen and possibly the corresponding lexicalised acoustic utterance (see fig. 1). These pairs are built by distributing all the same-length syntactic references of the corpus with each reiterant stimulus. The listeners' answers consist in either the reiterant stimulus association with the syntactic reference, or the dissociation of the two stimuli. Association and dissociation scores are the basic information used to achieve the results.

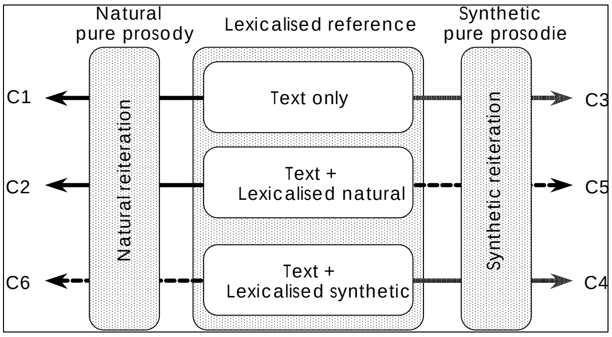

Figure 1: the six experimental conditions defined for the reiteration experiment, and the nature of the stimuli presented to listeners

3. Results analysis and comparison

3.1. Online subjective validation

The analysis of subjective results show no significant difference between the two synthesizers (see [15]). On the other hand, the different passages do not receive the same agreement from listeners: some structures seem to be more problematic, especially the combination of interrogative forms plus complex syntactic structures. The correlation between the number of word underlined by listeners and the global quality rating given to each passage is highly significant (r=.85, p>0.001).

Despite this good subjective result, the comparison with objective analysis does not allow the use of an acoustic parameter as a good predictor of the listeners' choices, neither for a local analysis, nor for a global rating. It is however interesting to note that the only acoustic parameter that matches sometimes the subjective results is - surprisingly - the duration one. That leads us to question the efficiency of the acoustic analysis, as it is not in accordance with classical descriptions of French prosody and the demarcation function.

3.2. Stimuli with incoherent prosody

At a first glance, there is no major difference in this experiment between the performances of synthetic and natural prosody. This first result underlines the good overall performances of the prosody generator [13].

However, a more detailed analysis allows a very precise diagnostic of the compared performances of synthetic vs. natural prosody. Thus, natural prosody is more "permissive" than the synthetic one, and one can miss some syntactic boundaries (between groups of 1 or 2 syllables), without producing an ill-formed utterance, whereas synthetic prosody with lacking boundaries is systematically noted by listeners. On the other hand, missing boundaries for longer groups, or added boundaries are perceived by subjects.

A problem raised by such a paradigm is the result obtained by synthetic prosody, comparable to the natural performances: is it really due to the high performances of generator, or is it a bias of the pair-presentation? If the two members of a pair are quite correct (even if they have not the same performances) the listener will not reject the last one, but rather accept both stimuli; leading to a leveling down to the synthetic performances.

3.3. Reiterant speech

Global performances for this experiment points out the minor importance of the experimental condition, and then the minor importance of the origin of the prosody (either natural or synthetic) - re-validating the quality of the generator. This parallel evaluation of natural and synthetic prosody can be compared. It shows that the relative performance of both prosodies, if they are globally equivalent, are (i) sometimes better for synthetic prosody (for simple syntactic structures), and sometimes better for the natural one.

The major variance in the results is induced by the syntactic structure of stimuli. An important result, as it shows the ability of listeners to perform a meta-linguistic task on the basis of reiterant prosody only: their association and dissociation scores are given accordingly to the classical description of the segmentation / hierarchisation function of prosody in French (cf. [8]; [4]).

Moreover, the analysis lead to the construction of a scale rating the perceptive divergence existing between two demarcation functions (see figure 2, which schematizes this divergence).

This compared analysis of the subjective vs. objective divergences between stimuli gives a detailed map of the prosody generator strength and weaknesses:

- Simple syntactic structures are very efficiently performed, maybe overlearned (synthetic > natural prosody).
- More complex syntactic structure (long clauses, enumeration) are deficient (natural > synthetic prosody).
- For the other syntactic structure, natural and synthetic performances are similar.
- Fundamental frequency map efficiently the perceptive results, for both natural and synthetic prosody
- Duration patterns are also very efficient predictors of perceptive results for natural prosody, whereas synthetic duration misses such a performance.

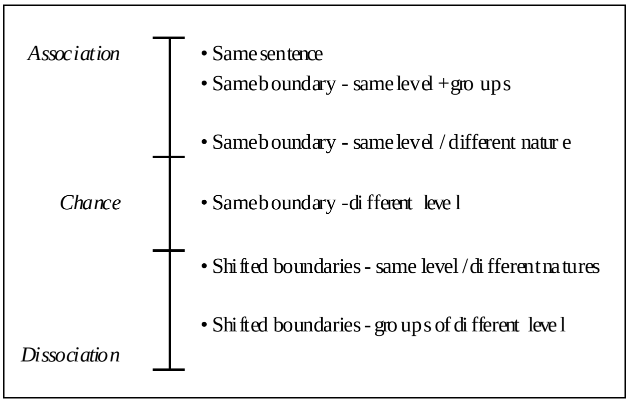
Association
Chance
Dissociation
• Same sentence
• Same boundary - same level + groups
• Same boundary - same level / different nature
• Same boundary - different level
• Shifted boundaries - same level / different natures
• Shifted boundaries - groups of different level

Figure 2: schematic scale representing the perceptive divergence between two segmentation / hierarchization functions.

4. Conclusions

By comparing the results extracted from each of these three experiments, it stands out that:

- The first experiment, even if it stays out of the actual metalinguistic process, induced by the decoding of the demarcation function allows to spot the major prosodic deficiency. Results can be interpreted by the developers of the prosody generator, as they are localized, and matched with an acoustic analysis, but they cannot be directly interpreted, as they are not sufficiently precise.
- The experiment using incoherent prosody is more directed to the competencies explicitly manipulated by the synthesizer (the demarcation function), in comparison with natural reference. But as the presence of this reference levels down the listeners' answers, it is more a validation of synthetic competencies alone, than a comparison with natural ones.
- The last experiment aims at directly and explicitly proposing to listeners to rate the realization of a function in a pure prosodic form. As natural and synthetic stimuli are evaluated in exactly the same way, but in two separate conditions, their performances can be compared and then the synthetic competencies can be diagnosed in a more efficient way. Moreover, as listeners manage to perform this task, it is a direct way to test the effective efficiency of prosody in performing a given function.

The combination of these three experiments draws up a picture of the actual linguistic abilities of the prosodic model in segmenting and hierarchizing speech, and allows to sort out the structure already learned, and the one that could benefit from a new learning round. Acoustic parameters are also tested, with good results for the fundamental frequency, and some problems for duration.

If we focus more on the cognitive processing of prosody, these experiments raise some problems: (i) in the pre-test made on reiterant speech, sentences longer than 11 syllables were largely rejected by listeners. On a similar experiment, [17], shows a degradation of listeners' performances after the same length; and (ii) the low precision of diagnostic performed in the first experiment could be due to the length of text passage

All these results question the ability of listeners to perceive prosodic information and to perform a metalinguistic processing from this basis on stimuli longer than 11 syllables. Such a result can be compared with hypotheses made by [6] on a specific processing of prosody, limited to a similar length – that should be a specialization of the articulatory loop.

An extension of these experiments could be raised from these results. First, to test the ability of a listener to perform direct online diagnostic on shorter stimuli (the first experiment with the stimuli of the third one). The second paradigm must be enhanced by using a direct ranking of each perturbed stimulus, instead of a compared one.

5. Acknowledgements

We are deeply grateful to Daniel Hirst, Gérard Bailly and Yann Morlec for their contribution in those works and their fruitful advice.

6. References

[1] Aubergé, V., 1991. *La synthèse de la parole : des règles au lexique*. Thèse de doctorat en informatique. 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. *23rd JEP*. Aussois, France, 281-284.

[3] Beaugendre, F, 1994. *Une étude perceptive de l'intonation du français. Développement d'un modèle et d'une application à la génération automatique de l'intonation pour un système de synthèse à partir du texte*. Thèse de 3ème Cycle. Univ. Paris XI, Paris, France.

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

[5] Campbell, N., 1998. Where is the information in speech? *3rd ESCA/COCOSDA Workshop on Speech Synthesis*. Jenolan Caves, Australia, 17-20.

[6] Gérard, C.; Dolgër, N., 1996. Taille des fenêtres perceptives, empan de la mémoire auditive. *21th JEP*. Avignon, France, 59-62.

[7] Hirst, D.J.; Nicolas, P.; Espesser, R., 1991. Coding the F0 of a continuous text in French: an experimental approach. *XIIth ICPhS*. Aix-en-Provence, 5, 234-237.

[8] Hirst, D.; Di Cristo, A., (Eds.) 1998. *Intonation systems: a survey of twenty languages.* Cambridge University Press.

[9] Hirst, D.J.; Rilliard, A.; Aubergé, V., 1998. Comparison of subjective evaluation and an objective evaluation metric for prosody in text-to-speech synthesis. *3rd ESCA/COCOSDA Workshop on Speech Synthesis.* Jenolan Caves, Australia, 1-4.

[10] Larkey, L.S., 1983. Reiterant speech: an acoustic and perceptual validation. *JASA* 73(4), 1337-1345.

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

[12] Mertens, P.; Auchlin, A.; Goldman, J.P.; Grobet, A., 2001. L'intonation du discours : une implémentation par balises ; motifs et premiers résultats.. *Journées Prosodie 2001*, Grenoble, France.

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

[14] Morlec, Y.; Rilliard, A.; Bailly, G.; Aubergé, V., 1998. Evaluating the adequacy of synthetic prosody in signalling syntactic boundaries: methodology and first results. *1st LREC*. Granada, Spain, 647-650.

[15] 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.

[16] Rilliard, A.; Aubergé, V., 1998. Reiterant Speech for the Evaluation of Natural vs. Synthetic Prosody. *ICSLP'98*, Sydney, Australia, 675-678.

[17] 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.

[18] Vannier, G.; Lacheret-Dujour, A.; Vergne, J., 1999. Pauses location and duration calculated with syntactic dependencies and textual considerations for t.t.s. system. *ICPhS 1999.* San Francisco, USA.

[19] Van Santen, J.P.H., 1997. Prosodic modelling in Text-to-Speech synthesis. *EuroSpeech'97*. Rhodos, Greece, KN 19-28.