PROSODY DIAGNOSTIC USING REITERANT SPEECH

Albert Rilliard & Véronique Aubergé
Institut de la Communication Parlée, Université Stendhal/INPG
Domaine Universitaire - 1180 Av. Centrale - BP 25 - 38040 Grenoble Cedex 9 - France
{rilliard, auberge}@icp.inpg.fr

ABSTRACT
This paper describes a set of experiments using French reiterant speech (on a canonical [ma] syllable). Experiments are designed to perform a diagnostic evaluation of the linguistic performances of synthetic prosody. Different experimental procedures are organised to match either synthetic or natural utterance, in their reiterant or lexicalised versions. The natural procedures are used as a reference (natural prosody is supposed to be well-formed), synthetic ones test the performance of the synthetic prosody, and the mixed synthetic and natural conditions (presented here) measure the distance between the ideal natural prosody and the actual synthetic realisation. The Analysis of the results of the last two experimental conditions shows this paradigm ability to point out which linguistic function is performed by a given prosody, and into what extents.

1. INTRODUCTION
During the passed years, an increasing need for new paradigms to evaluate the quality of synthesised speech and more specifically to improve the diagnostic function of evaluation was claimed by scientists [1], [2]. Such diagnostic analysis of synthesiser are useful to improve the systems, or to test the capabilities of each module [3], [4]. Evaluation can be held in different ways: (i) by means of subjective quality/acceptability judgements; (ii) by adequacy rating of the synthetic utterances, made by a direct comparison with a natural reference (in a subjective and objective way); and (iii) by the evaluation of the linguistic competence, through the functional equivalence between a natural and a synthetic sentence. We present here a method testing the functional equivalence of a synthetic prosodic continuum vs. a natural one.

Constructing such a diagnostic of prosody alone requires to be able to deal with prosodic informations only, without interference from the other linguistic levels. There are at least two ways to neutralise the influence of the other linguistic levels: (i) counter-balancing the influence of the disturbing levels by, for instance, transplanting different sets of prosodic parameters on the same lexical structure (see [5]); or (ii) forbidding the access to any linguistic level except to the prosodic one, by using delexicalised speech. Delexicalised speech is useful for many purposes, and many tools have been proposed in literature: the use of filtered speech [6], “nonsense” speech [7] [8], the PURR method [9], or reiterant speech [10] [11].

Since a perceptive study, held by Larkey [12], showed the adequacy of reiterant speech (on a canonical [ma] syllable) to keep and carry pertinent prosodic informations for listeners; and because of its interesting acoustic properties (neutralisation of individual segment intrinsic duration, continuous voicing of the [ma] syllable), we chose this method to produce delexicalised speech.

2. METHODOLOGY
As explained above, the experiments presented here are based on the use of [ma] reiterant speech, and were designed to test the adequacy of a synthesised prosodic continuum for a linguistic function. We are here interested in one of the function held by prosody, that is the hierarchisation and segmentation function, as it is of primary importance to provide a well-formed synthesised speech.

The experiments use a set of declarative read French sentences, extracted from a corpus (built to design the ICP prosodic model [13]), and respecting a set of minimal pair of syntactic oppositions. Each sentence is produced in a synthetic or a natural version, in its lexicalised and reiterant form.

2.1. Theoretical Design
Experiments are based on the systematic opposition of a lexicalised vs. a reiterant sentence, the synthetic and natural version mixed through the different experimental conditions. It results in six different experimental conditions, differing in the nature of the stimuli proposed to listeners, but all based on the same experimental procedure; each single condition tests a precise kind of information. Experimental conditions are summarised in table 1.

<table>
<thead>
<tr>
<th>Condition number</th>
<th>Reiterant stimulus</th>
<th>Lexicalised stimulus</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>natural reiteration</td>
<td>text alone</td>
</tr>
<tr>
<td>C2</td>
<td>natural reiteration</td>
<td>text + natural speech</td>
</tr>
<tr>
<td>C3</td>
<td>synthetic reiteration</td>
<td>text alone</td>
</tr>
<tr>
<td>C4</td>
<td>synthetic reiteration</td>
<td>text + synthetic speech</td>
</tr>
<tr>
<td>C5</td>
<td>synthetic reiteration</td>
<td>text + natural speech</td>
</tr>
<tr>
<td>C6</td>
<td>natural reiteration</td>
<td>text + synthetic speech</td>
</tr>
</tbody>
</table>

Table 1: nature of stimuli for each of the 6 experimental conditions.

The C1&C2 conditions (conditions presented in [14], the two plain natural conditions, are designed first to test the feasibility of this paradigm (the use of reiterant speech, as a non-ecological material requiring a metalinguistic ability, has to be validated), and secondly to serve as a reference to the next conditions. As natural stimuli are supposed to be well-formed, the question raised by the C1 condition which is “how far” can prosody facilitate the syntactic segmentation and hierarchisation function. The C2 condition, as it proposed an oral lexicalised reference to listeners, answers the topic of the “linguistic intelligibility” of natural prosody.
Conditions C3&4 (see [15] for a more detailed analysis) are the plain synthetic conditions, and propose an evaluation of synthetic prosody. More specifically, the result of C3 can be directly mapped to C1, to point out the major differences between the natural and the synthetic function efficiency. The interpretation of C4 results is more problematic: in the C2 condition, subjects use the natural lexicalised sentence as a well-formed reference to judge the functional adequacy of the reiterant prosody; but the lexicalised sentence in C4 is a synthesised one, and we are not aware of its well formedness. Then, to interpret the listeners’ answer, we need a quality rating of the lexicalised synthetic sentence, that can be retrieved in the C5 condition. C4 was performed anyway, because of the poor quality of synthetic stimuli (in comparison to the natural ones), that only perform the segmentation and hierarchisation function: their computation by listeners should be more simple.

Results of the C6 condition have to be compared to C5 results. As both conditions are orthogonal, if the results are consistent, we will conclude that the synthetic stimuli presented in the C6 condition carry an adequate variant of the C5 natural one, for the same function. In the contrary case (non-concordant results), the paradigm itself should be reconsidered, as the subjects do not give the same answer for both conditions.

Table 2: some representative examples of syntactic construction from the corpus

<table>
<thead>
<tr>
<th>Enumeration</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Je mangeais du vin, du Boursin, et du pain.&quot;</td>
</tr>
<tr>
<td>&quot;I was eating wine, Boursin and bread&quot;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Adjective/Noun Opposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Ce beau passant chantait.&quot; vs. &quot;Ce passant fou chantait.&quot;</td>
</tr>
<tr>
<td>&quot;This beautiful passer-by was singing.&quot; vs. &quot;This crazy passer-by was singing.&quot;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NG-VG vs. Clause–Clause Opposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Ce passant chantait tous les six mois.&quot; vs. &quot;Ce passant chantait, Toto dansait.&quot;</td>
</tr>
<tr>
<td>&quot;This passer-by was singing every six months.&quot; vs. &quot;This passer-by was singing, Toto was dancing.&quot;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NG-VG vs. GV-GO Opposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Ce beau passant chantait.&quot; vs. &quot;On entendait des pas.&quot;</td>
</tr>
<tr>
<td>&quot;This beautiful passer-by was singing.&quot; vs. &quot;We heard some steps.&quot;</td>
</tr>
</tbody>
</table>

The possible answers for the instructions are “Yes” or “No”. Subjects also have to give the confidence level they feel towards their answer, on a 1 (quite sure) to 5 (not sure at all) scale. Then, they click onto a button to start the next stimuli. The reaction time between the end of the sound presentation and the last click is recorded for each stimulus presentation.

The stimuli are presented at random, without the same sentence being presented twice successively (either the reiterant or the lexicalised ones). A given stimulus is presented only once to each listener.

A sentence will appear on the screen. Please read it carefully and press the button “Listen”.

You will hear a sentence in which each syllable is replaced by the syllable [ma], and then hear the sentence written on the screen, normally pronounced. Both sentences have the same number of syllables. You will have to answer “yes” or “no” to the following question: “Could the mamama utterance you have just heard be rightly associated to the sentence on the screen?”

Next, you will have to give your level of confidence when answering the question, on a scale ranging from 1 (“I am positive about my answer”) to 5 (“I feel very uncomfortable about my answer”). Then press the button “OK” to switch to the next question.

2.2. Experimental Design

2.2.1. Procedure. The basic experimental procedure, underlying each condition, is identical. It is based on the successive presentation of stimuli couples, by way of a computer screen for text display and headphones for the recorded sentences. The stimuli couples construction follows the condition description summarised in table 1.

For each stimuli couple, the text of the sentence is displayed first on a computer screen. The subject reads it and clicks onto a button to hear the related sentence. Reiterant then lexicalised (if the condition requires it) sentences are played once, in headphones, at a comfortable hearing level. The instructions given to listeners are translated in figure 1.

We should note that the paradigm used for the C1&2 conditions presents some differences from the paradigm described here. This is the consequence of the exploratory character of these two conditions. They were held first, during a first exploration of the reiteration paradigm. The new paradigm used for the next experiments is the result of what we have learnt from these first conditions.

The presentation of the stimuli is the same as in the new paradigm. Major differences can be found in the results format: listeners have to answer the same question (“Is the reiterant prosody adequate for the sentence presented on screen?”) by “Yes, “No”, or “I don’t know”, instead of a simple Yes/No answer plus a confidence scale. Reaction times were not recorded for these first two conditions.

2.2.2. Stimuli. As already said, the corpus sentences are based on a corpus made for the construction of the ICP prosodic model
analysis-resynthesis method (using a TDPSOLA coder). The synthetic reiterations) sentences. Then the prosodic parameters for each sentence are applied on these flat recordings by an analysis-resynthesis method (using a TDPSOLA coder). The parameters for the synthetic stimuli (either lexicalised or reiterant) are taken from the output of the ICP synthesiser; and parameters for the reiterant natural stimuli are calculated on the basis of the lexicalised natural stimuli. Such a method was used to avoid the low segmental quality of the synthesiser’s output, which can influence the listener’s judgement.

2.2.3. Subjects. The subjects of these experiments are all native French speakers, without any listening problem. 13 listeners performed the C5 condition, and 7 the C6 one. The C6 condition is not yet finished and more subjects should performed it.

3. RESULTS
The results concerning the C1&2 conditions were presented in [14], and the results concerning the C3&4 conditions in [15]. We are interested in this paper in the analysis of the C5&6 conditions results, and because they need to be interpreted under the light of the C2 results, we will shortly introduce them. Then, results from the C5&6 condition will be described and analysed.

3.1. C2 Condition Summary
This condition proposes to listeners a reiterant natural sentence matched with text and a natural lexicalised sentence. Association scores (“Yes” answer to the question “Is the prosody adequate?”) are high for homogeneous stimuli (both reiterant and lexicalised sentences based on the same sentence). This result is in accordance to the validation procedure of the natural stimuli quality [14]. For the heterogeneous stimuli (the reiterant sentence divergent from the lexicalised one), results are in accordance to the classical descriptions of prosody for its segmentation and hierarchisation function (see [16] for example): emergence of major syntactic groups, primary importance of the syntactic boundary placement, and lower influence of the level and nature of the syntactic components.

3.2. The C5 condition
The homogeneous stimuli (direct validation of the prosodic adequacy, compared to the natural reference) present a low association score (73% vs. 90% in C2). Such a first result illustrates the somewhat poor overall quality of synthetic prosody, in comparison to the natural complexity.

A more detailed analysis points out some sentences as badly produced by the synthesizer. Two structures receive a 70% dissociation score, and two others are around the average. One of the badly rated sentence has an enumeration structure, which is not a common structure in the building corpus of our prosodic model. The three other structures are clauses, more difficult to produce.

Heterogeneous stimuli, compared to the C2 results expose some differences in the subject’s answers. Such differences can be divided into two categories:

First the unrecognised prosodic shapes. That is the prosodic parameters, validated in the homogeneous comparison, but unable to discriminate other structures, whereas natural prosody is able to discriminate it. We found some examples of such weaknesses for the segmentation and the hierarchisation functions.

Secondly the adequate prosodic variant: a badly produced sentence is found to be adequate for another structure (that natural prosody does not associate). The synthesiser has involuntarily produced a variant of the second structure. The example of this behaviour is the enumeration sturcture reported on a clause structure.

3.3. The C6 condition
The C6 condition proposes to listeners a natural reiteration associated with a synthesised lexicalised “reference”. Results need to be interpreted by comparison to C5 ones.

For the homogeneous stimuli, results show a specific behaviour, in respects to the other conditions: the synthetic utterances known to be incorrectly produced are rejected around the average, but some other sentences are rejected by listeners, that were never extracted by the other conditions.

For the heterogeneous ones, major tendencies are respected but the overall results are brought back closer to the average.

4. ANALYSIS
The result analysis is based on a systematic comparison based on syntactic oppositions, of the different conditions. In a first step, the homogeneous stimuli, then each major oppositions listed in the C2 condition description are reviewed. All the results can not be describes hereafter in extension, but we will try to explain and summarise the kind of information that can be extracted from such an analysis.

4.1. Same sentences
Homogeneous comparisons allow a direct validation of the relevance of the synthetic prosody for the task it was produced. It is the basic evaluation level of our paradigm, with a binary answer: is this prosody good or not?

4.2. Different sentences
Heterogeneous oppositions are the diagnostic part of our experiment. Such oppositions are designed to test, for each minimal pair oppositions if (i) the tested prosodic parameters are a possible and acceptable variant - and the acceptability level of this variant; (ii) the relative contribution of prosody in the realisation of a given linguistic function; and (iii) the distance (in terms of performance) between the natural reference and the tested synthetic prosody.

For example, the assimilation by listeners of two sentences opposed by the hierarchical level of their syntactic component exhibits the lack of our model to produce such a distinction.

The major oppositions tested during our experiments are (i) the localisation of the major syntactic boundary, which is not sensitive enough in regards to the natural prosody performances; (ii) the nature of the sentences major syntactic groups, which seems to be a low informative indices for natural prosody, and
completely insufficient for the synthetic one; and (ii) the hierarchisation indices, that is not relevant for our synthetic speech.

We note that both the hierarchisation and the nature indices in combination generate a pertinent discrimination indices for listeners; that can be interpreted as the summation of two insufficient indices: the synthesiser seems to produce hierarchisation and nature indices, but a single one is too weak to be adequately perceived by listeners.

5. CONCLUSION
The paradigm proposed here is not simple to construct and to drive out. Some questions can be raised about the task, that is a non ecological one, required the explicit use of metalinguistic abilities from listeners; about the adequacy of reiterant speech as a tool for evaluating linguistic material, and as a material that can be produced by synthesisers. The problem of the listeners’ ability to make the most of metalinguistic informations on the basis of reiterant speech has been raised, as soon as the listeners’ answer for natural prosody is coherent with the known functions of prosody, and answers for synthetic speech extract really bad formed sentences.

Moreover, the strength of such a paradigm is its ability to produce a very precise and tuneable diagnostic of prosody, with a complete decorrelation from the underlying linguistic material. It fulfils our attempt to propose an alternative paradigm that allows (i) an analysis of the performance of a given synthetic prosody with reference to a natural reference; and (ii) to diagnose the competence of the model which produced the tested prosody. Each different experimental condition from C3 to C5 behaves like a series of filters, with a different selectivity: C3 selects only the worst sentences; C4 has a central position; and C5 is very selective, as the appropriateness of synthetic prosody is matched to a natural reference.

ACKNOWLEDGMENTS
Many thanks to Gérard Bailly and Yann Morlec for their great help in this research and in the construction of the stimuli.

REFERENCES