The Third ESCA/COCOSDA Workshop
(ETRW) on Speech Synthesis
Jenolan Caves House, Blue Mountains,
NSW, Australia, November 26-29, 1998

Reiterant Speech for the Evaluation of Natural vs. Synthetic Prosody

Rilliard Albert & Aubergé Véronique

Institut de la Communication Parlée – Grenoble, France

email : {rilliard, auberge}@icp.inpg.fr

ABSTRACT

This work deals with some evaluation experiments on reiterant speech using both synthetic and natural stimuli. They have been designed to test the efficiency of the described paradigm to diagnose the adequacy of synthetic prosody to syntactic structure in reference with natural performances. Following a general methodology developed for synthesis [7], experiments have been conducted on the ICP synthetic prosody with the aim of validating the original corpus from which the prosodic model was learned .

1. INTRODUCTION

Prosody performs many linguistic functions (enunciation structuration – segmentation and hierarchisation of sentences, dialogue, discourse – modalities, focalisation) and extra-linguistic functions (attitudes, emotions). Our long-term goal is to be able to identify and to measure the competences, through its performances, involved by prosody in the completion of a given communication task. A grid could then be built to list, for each linguistic function, the respective importance of prosody in its realisation depending on the situations and on the speakers' strategies. Such a grid for natural prosody could be a reference for similar grids for synthetic prosodies.

This is the general perspective; the first step is the different methods presented here which deal with the adequacy of synthetic prosody to perform a lexical and syntactic segmentation facilitation which is essential in the field of TTS, since these performances can be related to the "linguistic intelligibility" of speech. This work is dedicated to the study of French declarative utterances, for which prosody is supposed to be coherent with the underlying syntactic structure.

How to measure the adequation of prosody *alone* during the segmentation and hierarchisation mechanisms; without any interaction of the other linguistic agents performing the same functions, with or without any redundancies? A first solution is to *neutralise* the effect of (i) the syntax (see for example preference tests on well-formed vs. bad-formed prosodic utterances in [7]); or (ii) the semantic like in SUS test [2]. A second solution is to *suppress* the lexical information, that is all the upper structures, using delexicalised speech. This can be obtained by several ways, either by degrading the segmental quality of speech [3], or using nonsense speech [10], or by producing reiterant speech on a canonical syllable [8] [6]. We chose to use the reiterant speech paradigm, because of its interesting properties of normalisation [6], and of its conservation of pertinent information about the syntactic segmentation [5]. This allows a simplified measure of both the acoustic and perceptive information carried by prosody.

A preceding experiment already used such a kind of methods on natural sentences, through a validation test [9]. We present here the complementary experiments held with the synthetic prosody of ICP TTS system, since our short-term aim is also to evaluate the adequacy between the synthetic prosody and the natural prosody of the original corpus from which the prosodic model was learned [7].

We describe after the general different ways in which such an experiment can be held with varying aims. The firsts results of experiments using synthetic speech are also presented.

2. PERCEPTION EXPERIMENTS

2.1. Different conditions for different evaluations

The principle of the tests is to submit to listeners a pair of stimuli whom adequacy/similarity they have to decide. The stimuli are built in order to represent syntactic boundaries varying on the syntagmatic and paradigmatic axis. The first (tested) stimulus of the pair is reiterant speech, the second one is either a written sentence or a written sentence with a speech stimulus. The interest of this method is to give a direct access to syntactic association/dissociation competences, in comparison with some indirect methods, as for example [7]. As a correlate, the main problem of such experiments is in being a complex cognitive task, since it is quite a metalinguistic condition. As usual in such tasks, nothing can guarantee that the subject is able to extract the required information as sub-products of his cognitive processing, even if he really uses these information during his treatment.

All the possible combinations of pairs are presented to the listeners (a distribution of a reiterant stimulus with any stimulus of the corpus with the only constraint of having the same syllable-length). The stimuli are thus separated between homogeneous (the reiterant sentence matches its original syntactic construction) and heterogeneous ones. The experiments will then yield two kinds of linked results: association rating between homogeneous stimuli, and dissociation rating for heterogeneous stimuli. With such material, the possible presentation paradigms are listed on table 1. In previous experiments (see [9] for details), the tested (reiterant) stimulus was natural speech associated to text in C1 (table 1) or text and natural speech in C2 – which can involve both linguistic and acoustic associations.

Enumeration
"Je mangeais du vin du Boursin et du pain"
"I was eating some wine, Boursin and bread""

Adjective / Noun Opposition
"Ce beau passant chantait." vs. "Ce passant fou chantait."
"This passer-by was singing."
vs. "This crazy passer-by was singing."

NG-VG vs. Clause–Clause Opposition
“ Ce passant chantait tous les six mois. ”
vs. “ Ce passant chantait, Toto dansait. ”
"This passer-by was singing each six month."
vs. "This passer-by was singing, Toto was dancing."

NG-VG vs. GV-GO Opposition
“ Ce beau passant chantait. ”
vs. “ On entendait des pas. ”
"This beautifull passer-by was singing."
vs. "We heard some steps."

Figure 1: Some examples of sentences extracted from the corpus.

The natural prosody is, by definition, well-formed. Thus, for C1, the question raised by the test is "how far" can prosody facilitate the identification of syntactic structure. For C2, the question raised by the test is directly the degree of "linguistic intelligibility" of natural prosody, since two natural (reiterant and lexicalised) stimuli had to be associated.

On the contrary, the synthetic prosody cannot be supposed to be well-formed, since it is precisely one of the results expected from the tests. The test to be held on synthetic speech can not be reduced to an association between homogeneous stimuli, because our prosodic model is based on the principle of superimposed carried and carrying contours [1] – a motivation for the dissociation test in experiments C1&2 was precisely to retrieve some indices about this principle for natural prosody. This means that we need to test the strength of the contours' levels: we will see in the results that some different sub-level-contours will be identified as variants.

Thus, the same conditions as C1 and C2 must be applied for synthetic speech (C3 and C4). But they do not perform exactly the same actions. C3 results can answer directly if the prosody is well-formed in comparison with C1 results as a reference. C4 is more complex, since the lexicalised synthetic stimulus associated with the reiterant stimulus is not an a priori well-formed reference. Thus, even if C4 is supposed to test, as C1, the degree of linguistic intelligibility of (synthetic) prosody, the results will be filtered by the same bias, that is the possible bad-formedness of synthetic stimuli. C4 was performed anyway because the synthetic stimuli are poor and selective as compared to natural stimuli, since they perform only the segmentation/hierarchisation function.

C5, which associates synthetic reiterant stimuli with text and natural speech avoids the bias of C4 and cannot be compared to C2, since the well-formedness of the reiterant stimuli is tested. Moreover, the results can be interpreted only if they are related to the results of C4; when they are similar, the synthetic stimuli are well-formed, in the other cases the pairs accepted by the listeners contain acceptable prosodic variants, involuntarily produced by the synthesiser.

The last experiment C6 is orthogonal to C5. When the results of C6 and C5 are similar, it will mean that synthetic and natural are equivalent variants for the same function. In other cases, the test will evaluate the reiteration paradigm itself. The results of experiments C5 and C6 will be presented in a further paper.

Condition number	Reiterant Stimuli	Syntactic structure
C1	Natural reiteration	Text alone
C2	Natural reiteration	Text & Natural stimulus
C3	Synthetic reiteration	Text alone
C4	Synthetic reiteration	Text & Synthetic stimulus
C5	Synthetic reiteration	Text & Natural stimulus
C6	Natural reiteration	Text & Synthetic stimulus

Table 1: possible experimental conditions with synthesised or natural reiterant and synthetic stimuli.

With such material, the possible presentation paradigms are listed on table 1. We will describe hereafter the conclusions we can base on each kind of experimental condition.

The basic experimental condition is thus the association test between a prosodic sentence (that is the reiterant stimuli) on the one hand opposed to a lexical structure (proposed under one or two modalities) on the other hand. The reiterant stimulus is presented as an oral percept, and the lexical structure is presented as a text displayed on a computer screen with or without an additional oral presentation. Listeners only hear each signal once, and have to answer the question: "Is the reiterant sentence you heard compatible with the sentence displayed on the screen?", by "Yes" or "No". They also have to give a confidence level, on a scale from 1 (quite sure) to 5 (not sure at all). The reaction time (RT), for C3 to C6 only, is recorded also from the end of the oral presentation to the validation of the answer, in order to evaluate the cognitive load implied in each (diss-)association. Unfortunately, we did not measure the RT for C1&2, thus it will not be possible to compare processings involved in natural vs. synthetic stimulus for the same task.

The stimuli used for the experiment belong to a set of sentences extracted from the ICP synthesis corpus from which the prosodic model has been learnt [7]. Sentences were selected on the basis of a set of syntactic oppositions, based on the minimal pair principle. The same corpus was used for the validation test in order to diagnose which kinds of structure were well/bad learned in the model.

There are 22 sentences from 5 to 11 syllables length. They are simple sentences on the basis of nominal, verbal and object groups and some simple clauses combination, varying the location of the same segment for different functions (see some examples on figure 1). The original corpus has been produced by one speaker, in a sound-proof booth, and digitalised as 16 KHz and 16 bits AIFF sounds. For experiments C1&2, two speakers listened and repeated each original sentence, and subsequently produced a reiterant version [9].

For C3 to C6, the synthetic stimuli, lexicalised and reiterant, are produced with the ICP prosodic generator: first, a speaker records the natural recto-tono lexicalized and "mamama" stimuli; second, the synthetic stimuli are generated by joining the flat sentences with the duration and fundamental frequency from the prosodic model, using a TDPSOLA analysis-resynthesis technique. In the same way, the "natural" stimuli are generated in applying the original natural duration and fundamental frequency on the flat sentences. This procedure has been used to ensure a ceteris paribus condition.

As might be expected [4] C1&2, the natural sentences containing a silent pause were systematically well dissociated from stimuli with different or without any pause. But it has to be noted that in the association that pauses, in C1&2, Consequently, in order to focus on finer prosodic indices, the prosodic generator was forced to produced the stimuli without any silent pauses. For a few stimuli, this constraint was too strong and the model had to generalise duration out of its limits. It is clearly the case of the stimuli representing some enumerations, always learned with pauses in the corpus, inside a nominal group (see 1 in figure 1).

3. RESULTS

% Good answers	C1&2	C3	C4	C5
homo. rating	92	86	74	66
conf. level	-	1.6	1.7	2.3
mean RT	-	286	342	351
hetero. rating	55	56	46	51
conf. level	-	1.9	1.9	2.3
mean RT	-	436	337	332

Table 2: Global results for C1 to C5

Table 2 presents general results on all the conditions. Results for C3 and C5 for synthetic stimuli are detailed after.

3.1. Condition 3: Synthetic Reiteration vs. Text Alone

Homogeneous pairs. Association results are high: 86% (between 73 to 100% of association answer), with a mean confidence rating of 1.6. As a reference, the score for natural stimuli in C1&2 was significantly better: 92% of association. This difference can be partially explained with the help of a set of three sentences (the enumeration simple sentence, and two complex clausal structures), which are clearly confirmed as bad-formed in all the following tests. But more generally, some sentences, which are identified as well-formed in the following tests, receive, as expected, lower scores in synthetic presentation because they do not contain pauses.

A significant correlation is noted between the association score and the confidence level (0.81), and between the confidence level and RT (0.77).

Heterogeneous pairs. The way a reiterant stimulus is dissociated from a set of syntactic combinations (with no acoustic help, since the second stimulus is only text) gives directly some indices on the structurating relevant cues carried by prosody. On the contrary, the non dissociation performances give access to perceptively equivalent prosodic contours for different syntactic structures. The mean score of good answers (right dissociation) is 56% (compared to 55% for C1&2), with a mean confidence rating of 1.9. We will see in the following experiments (see table 2) that the dissociation score is quite stable for C3 to C6. It must noted for C3 that the well dissociated pairs get mainly high scores (55% of sentences are dissociated between 70% and 100%) and the not dissociated sentences have bad scores of dissociation, with the same mean confidence rating (30% of sentences are dissociated under 30%).

Morover, the RT for heterogeneous stimuli (436 ms) is globally 1.5 higher than for homogeneous stimuli (286 ms). No correlation can be noted between the dissociation scores neither with the RT, nor with the confidence rating.

3.2. Condition 4: Synthetic Reiteration vs. Text and Synthetic Stimuli

In this experiment, subjects have to associate or to dissociatethe reiterant stimulus with text and a synthetic stimulus, generated from the text.

Homogeneous pairs. Association results are lower than in C3, that is 74% with a mean confidence rating of 1.6. Results are spread on a larger scale than in C3: 55% of sentences are associated between 70% and 100%, 45% of sentences are associated between 70% and 38% (15% of sentences are "associated" under 40%).

Heterogeneous pairs. The mean score of right dissociation, 46% with a confidence level of 1.9, is not shared like in C3. As for the results of homogeneous pairs, they are distributed on a larger scale: 10% of sentences are between 70% and 100% of dissociation, 20% of sentences are under 30%, that means than 70% are between 70% and 30%.

For this experiment, contrary to C3, the RT for heterogeneous (342 ms) and homogeneous (338 ms) stimuli are similar and is not significantly different from the mean RT of C3.

The main difference between the results of C1 (natural in same conditions than C3) and C2 (natural in same conditions than C4) is the increasing of the subjects confidence . But the global scores of association and dissociation are similar, and the sets of well/bad-formed sentences are the same. We can see in the previous results for C3&4 than the addition of an acoustic synthetic reference to the text disturbs the results, as expected since the reference is not a priori well-formed. So C4 is a cross-confirmation of badly-formed (that is to say disturbing) indices.

3.3. About Reaction Time

We can say that subject are coherent. RT mean are all around 300 ms, excepted for the heterogeneous presentations of condition C3. The standard deviations of RTs are also very coherent, around 90 ms for all the presentations. The RTs are significantly correlated to the confidence level, for the homogeneous presentations in all conditions.

4. DISCUSSION

Concerning the homogenous pairs, we can see that the association results are C3>C4>C5 (in C1&2, each sentence is well associated, since a previous experiment was held before C1 to select the "best" natural reiterations [*], it is consequently meaningless to compare the homogenous pairs scores between c1&2 and the others conditions). More precisely, at the sentence level, it can be noted that the same set of sentences are associated in the synthetic conditions, but with a tuning selective filter from C5 down to C3, that is C5 is a more constraining condition which keep high scores only for the "best" sentences, C4 is a middle level where sentences are shared around the average, and C3 filters only the "worst" sentences.

These "worst" sentences are mainly enumerations and clausal sentences.

Concerning the non homogeneous pairs, to refer to natural in C1&2 is essential, since natural is the reference to identify which contours can be perceived as variants for a same structure – for different syntactic structures. That means that these undissociated syntactic structures are not carried by prosody.

The main difference with the natural results can be mainly explained by the fact that the synthetic speech carries, in the synthetic stimuli generated for these experiments, only the realisation, by our model, of segmentation/hierarchisation function, and nothing else which is obviously carried by natural speech. Moreover, the principle of our model, based on carrying and carried contours, forces this caricature: the results for synthetic are extremes in comparison to the natural.

The main results we expected from these experiments is to diagnose on which kind of structures the system has failed (in the sense of worse performances to the natural) and on which kind of structures it got similar scores that the natural, and even better scores because of a over-generalisation effect of synthesis.

4.1. Non Co-occurrent Boundaries

It concerns boundaries of clauses and groups. When clauses do not coincide, whatever their length, they are quite 100% dissociated, That is a better result than the natural sentences, even if pauses appear in natural speech and not in synthetic speech. When clauses coincide, but some groups inside do not coincide, the same set of sentences as in C1&2 are well dissociated, but some structures have lower scores, decreasing from C3 to C4.

4.2. Co-occurrent Boundaries

Same/Different Nature of Syntactic Group. The classification by C3&4 is the same than by C1&2., that is that the difference of the nature of the sub-groups do not imply a dissociation. This result for natural was for us an interesting possible confirmation of our global contour hypothesis. This hypothesis was tested in the prosodic generator. Thus, it is an interesting result, for our model, not to be able, like our original corpus, to discriminate such differences.

Same/Different Syntactic Level. Once again, the set of dissociated structures is the same as in C1&2, but the results are caricatural (for example, that is shared to extremes, for synthetic speech: when a clause boundary coincides with a group boundary, the pairs are better dissociated than in natural speech).

A few sentences, typical of specific structures, were identified to be differently associated/dissociated than in C1&2. We can now relate these structures to our modelization processing in order to improve it.

5. CONCLUSION

The set of experiments presented here as a general paradigm, and more precisely for two conditions on synthetic speech, are, clearly, rather complex experiments. First because the use of reiterant speech, even if it has been shown to be cognitively relevant [9], [5], is not an easy ecological material for listeners. Second, as we already mentioned, with such meta-linguistic tasks we cannot be entirely sure that the listeners retrieve, explicitely the expected information, even when they make use of this information during their cognitive processing. But the great advantage of such methods is to get the linguistic value of the extracted information directly when it is possible to extract it. Moreover, we think that the essential aim of evaluating synthetic prosody is (1) to compare the performance of a given sample of synthetic speech to the performance of a set of referential natural speech, previously evaluated as being ideal in a given situation (2) to diagnose the competence - that is the model efficiency - of the synthetic speech, in order to come back to improve the model after evaluating the system. We do not wish to claim, with this work, that reiterant speech is the best paradigm to perform these two tasks. What we would like to propose, however, with this series of experiments, is that there are some other alternative methods for evaluation (quite easier to control than for example SUS methods, which is in the same field) which could provide both an absolute scale for natural and synthetic speech, and which could compare directly the competence of natural and synthetic speech in a diagnostic fashion. (the results of these experiments, very briefly exposed in the discussion, is helpful to understand qualities and defaults of our prosodic model, and will be the base of new improvements).

6. ACKNOLEDGMENT

We thank Gérard Bailly and Yann Morlec for their great help and the time they spent in preparing the stimuli. This work was held with the help of COST 258 and AUPELF evaluation grant.

7. REFERENCES

1.Aubergé, V., “Developing a structured lexicon for synthesis of prosody”, In Bailly, G., Benoit, C., and Sawallis, T.R. (Eds.) *Talking Machines: Theories Model and Designs*, Elsevier Science, Amsterdam, p. 307-321, 1992.

2.Benoit, C., Grice, M. & Hazan, V., "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences", *Speech Communication, Vol. 18*, p. 381-392, 1996.

3.Carlson, R., Granström, B. & Klatt, D.H., "Some notes on the perception of temporal patterns in speech", *Proceedings of the IXth International Conference on Phonetic Sciences*, Copenhagen, Denmark, p. 260-267, 1979.

4.Grosjean, F., "Linguistic structures and performance structures: Studies in pause distribution". In Dechert, H. et Raupach, M. (Eds.) *Temporal variables in Speech: Studies in Honour of Frieda Goldman-Eisler*. The Hague: Mouton, 1980.

5.Larkey, L.S., "Reiterant speech: an acoustic and perceptual validation". *Journal of the Acoustical Society of America, Vol. 73 (4),* p. 1337-1345, 1983.

6.Liberman, M.Y. & Streeter, L.A., "Use of nonsense-syllable mimicry in the study of prosodic phenomena". *Journal of the Acoustical Society of America, Vol. 63 (1),* p. 231-233, 1978.

7.Morlec, Y., Rilliard, A., Bailly, G. & Aubergé, V., "Evaluating the adequacy of synthetic prosody in signalling syntactic boundaries: methodology and first results". *Proceedings of the 1st International Conference on Language Resources and Evaluation, Vol. 1*, p. 647-650,1998.

8.Oller, D. K., "The effect of position in utterance on speech segment duration in English", *Journal of the Acoustical Society of America, Vol. 54 (5),* p 1235-1247, 1973.

9.Rilliard, A. & Aubergé, V., "A Perceptive Measure of Pure Prosody Linguistic Functions with Reiterant Sentences", *Proceedings of the 5th International Conference on Spoken Language Processing*, Sydney, Australia, 1998, (to appear).

10. Strom, V. & Widera, C., "What's in the "pure" prosody?", *Proceedings of the International Conference on Spoken Language Processing, Vol. 3*, Philadelphia, USA, p. 1497-1500, 1996.