A PERCEPTIVE MEASURE OF PURE PROSODY LINGUISTIC FUNCTIONS WITH REITERANT SENTENCES

Rilliard Albert & Aubergé Véronique

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

ABSTRACT

We present here a perceptual measure experiment of the linguistic segmentation hints carried by prosody. The selected paradigm is a dissociation test between couples of stimuli. The sentences are made of several segmentation variations in group and clause levels, and couples are made on all possible combinations of two sentences from the corpus. 20 listeners are able to associate at the same time the similar area and syntactic level frontiers. They admit a single syllable translation on the position of the major syntactic boundary in the reiterated stimuli. They distinguish the couples that do not show the same frontiers. The results also show that listeners are puzzled (random choices) when the proposed frontiers delimit complex segments.

1. INTRODUCTION

One of the well known linguistic functions of prosody is the facilitation of the utterance segmentation by listeners. We suggest (following the same idea as [5]) that the linguistic functions of prosody can be measured with a grid matching linguistic functions and the respective importance of prosody in its realisation. The long-term aim of this work is to propose a method for building up such a grid.

Another derived short-term aim is the validation of a cognitive model of prosodic processing represented as a set of superposed hierarchical contours, each belonging to a different level of linguistic description, and communicating together on *rendezvous* points [1]. We tried to characterise the perceptive nature of such *rendezvous* points.

The very modest approach we present hereafter is limited to the study of prosody of read isolated French utterances, without any contextual directive to speakers. The sentences are extracted from the corpus used to establish the French model of prosody at the ICP laboratory [1]. In such a constrained task, we may pretend measuring effective prosodic variations for a segmentation task. We thus associate the quantitative measure of acoustic variations to the results of a perception experiment of segmentation in order to understand which morphologic bonds (that is prosodic morphology) carry the segmentation cues.

The effective measurement of the perceived prosodic variations, without any interaction with the other linguistic actors, is at stake. This problem will be addressed in a basic way by the suppression of access to lexical information. This delexicalisation can be performed in many ways. For example the degradation of the segmental quality [4], the use of glottograms [8], of nonsense speech [13], or the use of reiterant speech on a canonical syllable [ma] for [9] or several [ba], [ma], [da] and others for [11].

As we wanted to measure prosody both perceptively and acoustically, we used reiterant speech on the canonical [ma] syllable as a good way to preserve prosodic shape. The mimicry of a perceived sentence by a listener/speaker is a conservative way to produce utterances acoustically easy to analyse and carrying pertinent prosodic information for a perceptive use (see the studies of [9] and [7] for more details on the acoustical and perceptive transfer of prosodic information during the reiteration of sentences).

We describe hereafter the construction and validation of the reiterant speech corpus used for a perception experiment based on an association paradigm, designed to match pure prosodic sample to syntactic written and/or spoken sentences. Then we will see in what way such a method can be exploited to assess the quality of synthetic speech prosody.

2. EXPERIMENTS PROTOCOL

2.1. Corpus Constitution & Evaluation

Building up a corpus. We extracted 55 sentences from the corpus of our synthesis model, with the following criteria: the lengths of stimuli ranged from 4 to 13 syllables, and sentences were selected with respect of minimal pairs syntactic oppositions. The sentences were constructed on nominal, verbal or object groups and several clauses. Two speakers (a female and a male) recorded these sentences in a soundproof booth. Each speaker heard the original sentence aloud three times and repeated it once matching the original recording, and once replacing each syllable with a canonical [ma] syllable.

Validation protocol. The pairs of original and reiterant stimuli were presented to 10 listeners, with the sentence displayed on a computer screen. Listeners were allowed to hear them as many times as needed, to judge of reiterated sentence prosody quality. Judgements were made on a scale of 0 (reiterant prosody different from the original one) to 5 (reiterant prosody perceived as equivalent to the original one).

Results. The subject answers median for the whole corpus is 4. Major problems in the reiterant prosody were found for long sentences with ambiguous syllable constructions. Speakers seem to have problems in reproducing the adequate number of syllables of some specific constructions: for example, the utterance "avec un balai" (*with a sweeper*) is always reiterated with 4 [ma] syllables instead of 5. Despite

such a problem, it is hard work to produce adequate reiterant speech and a validation of the adequacy of reiterant prosody seems to be of prime importance before using the stimuli for other purposes.

From the original 55-sentence corpus, we select a 22-sentence one, with 5 to 11 syllable sentences and a reduced set of minimal syntactic pairs. All sentences in the final corpus are correctly rated by the subjects (median score of 5, and 4 for 2 sentences).

2.2. Dissociation Experiment

On the basis of the 22 utterance corpus, a dissociation experiment is performed. The aim of this experiment is to test the ability of listeners to match or separate a given prosodic phrase with a syntactic construction. We thus present to the listeners pairs of stimuli constructed from one reiterant sentence matching with any original syntactic construction in the corpus, having the same number of syllables.

These pairs of stimuli are made for the 2 speakers. Two required experimental conditions are: (i) with only the reiterant stimuli presented in an oral way; and (ii) with both the reiterant and the original stimuli presented in an oral way. For both conditions, the original reference sentence is displayed on the screen.

The 20 subjects (10 per condition) are asked to read the original sentence, to listen to the stimuli, and then to answer the question "*Does the reiterant sentence you heard match the sentence written on the screen?*" with "*Yes*", "*No*" or "*I don't know*". Listeners can only hear each pair of stimuli once. The whole corpus is presented twice randomly, both speakers alternated.

An acoustic analysis of the corpus was carried out in parallel, to extract FO and intensity information from the speech signal. The results of the perception experiment were analysed under the light of this acoustical information.

3. RESULTS, ANALYSIS & DISCUSSION

3.1. Global Analysis

The results of the dissociation experiment confirm the evaluation experiment: the association of a reiterant sentence with its original utterance gives high association scores: an average of 92% (by "association", we mean that listeners answer that the reiterant sentence matches the original sentence, and by "dissociation", the reverse).

Major prosodic indices for listeners correspond to a set of syntactic factors: position of the boundary between two syntactic groups, hierarchical importance of opposed groups (e.g. Nominal Group (NG) vs. Clauses), or changes of syntactic nature of compared groups (e.g. N.G. vs. Verbal Group (V.G.)).

Global results show that the size of the syntactic groups seems to affect the accuracy of the dissociation by listeners. The longer the group, the more listeners detect boundaries shift, or inner boundaries. For a group shorter than 4 syllables, inside syntactic boundaries do not seem to have a noticeable importance, whereas for longer groups, subjects detect the inside syntactic construction, and are able to dissociate differences of construction that do not seem to have any importance for shorter groups. This importance of the size of the syntactic groups is coherent with our analysis of prosody as a superposition of hierarchical contours.

Inside the corpus, 6 sentences were produced with a silent pause. Each couple of stimuli constructed with those sentences has very good dissociation or association scores (73% of dissociation for non co-occurent pauses), as we can expect from the work of [6].

The results for the two speakers seem to be equivalent, with a correlation of 0.89 between association answers. The major difference between the two experimental conditions is the rate of answer "*I don't know*", more frequent for the oral reiterant sentence alone (225 vs. 75 answers), but this just slightly modifies the proportions of answers (r=0.69 for the "*Yes*" answer, and 0.64 for the "*No*" answer). The average rate of answer "*I don't know*" is below 10%, with a maximum of 11%, when the two other choices are balanced - haphazard choices.

We will detail hereafter the answers of listeners for each kind of heterogeneous stimuli pairs (made of two sentences with different syntactic constructions), based on the major boundary position in each stimulus and on the nature of the two groups separated by the boundary.

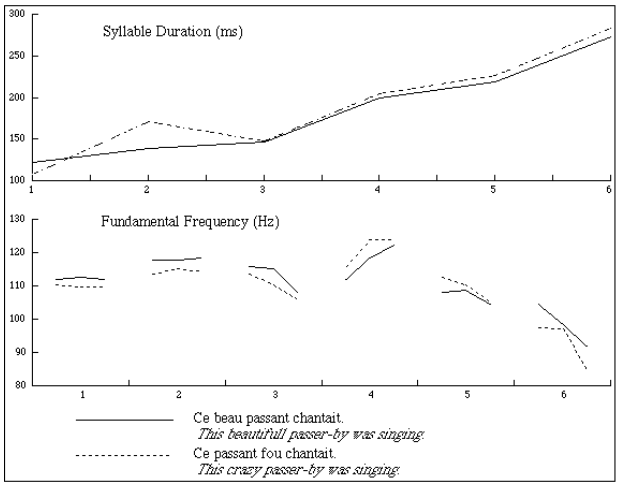

Figure 1: Adjective/Noun inversion in the Nominal Group. F0 and duration patterns are similar.

3.2. Co-occurent Boundaries

These kinds of comparisons are interesting insofar as they allow tests about the perceptive importance of prosodic

indices such as the nature of the syntactic groups, or their hierarchical level. Studies thus cannot be easily carried out if other indices perturb the prosodic continuum.

Syntactic groups of same nature. Such pairs of stimuli received the best association scores. From our corpus, these pairs are only represented by sentences composed of a NG/VG structure. 87% of the listeners associate these sentences. Differences in the constitution of the two major syntactic groups do not introduce significant indices for listeners, even the inversion of Adjective/Noun. Though, as the size of internal syntactic components grows, the dissociation scores increase. We interpret such a behaviour in the terms of our prosodic model as a perturbation of the carrier level interpretation by the information of carried level (cf. Fig 1) - see [1] for details**.**

Syntactic groups of different nature and same hierarchical level. For such a pair comparison, the association are, in average, of 51% (and 38% of dissociation). Such results do not show a clear distinction of the two prosodic contours. Those different results between presentation of stimuli pairs with the same groups, and pairs with different groups show that the prosodic continuum carries information about the nature of the syntactic group, but that it is a somewhat weak cue.

Syntactic groups of different natures and different hierarchical levels. Such comparisons are more perceptually different: 37% association vs. 52% dissociation. The addition of dissimilarities between compared segments finally produced a relevant cue to dissociate the different contours.

Emergence of perceptible sub-groups. Another interesting point that rizes from this kind of stimuli is the emergence of a significant influence of inferior syntactic description levels. From the different stimuli proposed to listeners, a set has comparable syntactic structure, but does not receive comparable discrimination results. An explanation that can be proposed is the size of the inner groups that is substantially larger in the better-discriminated pair of stimuli than in the associated one (4/6 vs. 3/2 syllables stimuli).

3.3. Non Co-occurent Boundaries

The major information that can be retrieved from these comparisons, concerns (i) the perceptual influence of the boundary shift length; (ii) the influence of errors on the listener's decisions; i.e. if a prosodic sentence fails to mark a syntactic index, can subjects reinterpret it in any other way to fit the proposed syntactic construction?

Perceptual influence of the gap between compared boundaries. We can note that a difference of syllable is a pertinent point to discriminate stimuli. The number of syllables discriminating the two boundaries increases the dissociation strength: a single syllable shift corresponds to 40% of dissociation; a 2 syllable one to 60% for comparable pairs of stimuli (7 or 8 syllable length, major groups of same syntactic nature).

The problem with such comparisons is the presence of silent pauses in a few signals, which are perturbing indices for the study of the accuracy of discrimination. For unperturbed signals, we can note the following behaviour: a one-syllable shift does not give sufficient information for a dissociation judgement. We refer to the fuzzy boundary notion introduced by [3] to describe such a phenomenon.

Missing indices. The pairs of stimuli described in the preceeding section are all made of somewhat simple sentences comparisons. We mean sentences composed of 2 major syntactic groups, each group being constructed on a basic structure. For stimuli made with more complex syntactic structures (or stimuli where the compared sentences are more different), a detailed analysis can be simplified if we consider this presentation as a combination of the simple cases listed before. For a simpler explanation of our purpose, we will shortly describe an example (cf. Fig. 2).

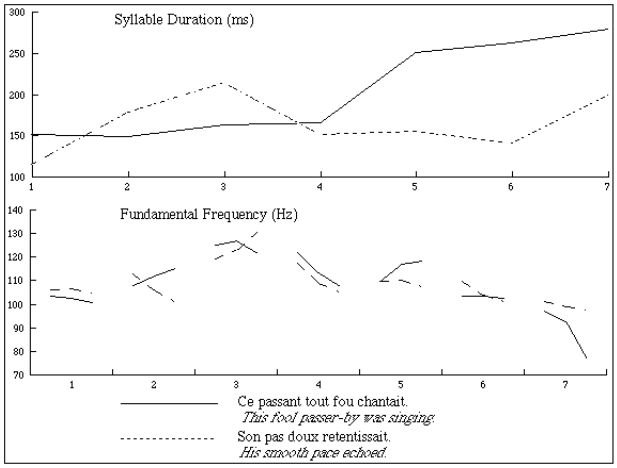

Figure 2: pair of stimuli with various indices.

Here, both sentences are composed of 2 major syntactic groups, on a NG/VG structure. The boundaries of these groups do not match. But, for the first matching, the major syntactic boundary of the reiterant sentence (dashed line) is at the same location as an internal boundary on the NG of the plain syntactic original sentence proposed (plain line). When perceiving the reiterant stimuli, subjects can detect an effective prosodic mark for the minor syntactic boundary, but do not perceive prosodic indices for the major one. Such a combination of an adequate prosodic hint with a lack of prosodic mark results in 40% dissociation and 40% association scores. The reverse presentation, where subjects heard an adequate key on the major syntactic boundary and an inadequate prosodic variation in the middle of the verbal group is dissociated by 70% of listeners.

As we can see from this example, listeners react differently to a false index or a missing index. A speaker can tend to speak faster, or try not to produce a perfectly correct sentence, by forgiving some index, but one can hardly allow someone to

add some prosodic variations (except for other linguistic purposes, as for a focalisation).

3.4. Other Results

Duration and speech rate measurements were carried out for the reiterant [ma] syllables. The perceived boundaries match with ascendant speech rate speed. This matching seems much more precise if we use the length of the Group Inter P-Centre (GIPC) (see [2] for details).

We used different duration units to analyse the result of the experiment: syllable, GIPC, and syllable *free* from their intrinsic and co-intrinsic duration, to obtain *pure* duration for the original and reiterant syllables (see [1] for a definition). The method that best fits the result of the perception experiment is that of the GIPC duration modelling.

4. CONCLUSIONS

This first experiment of perceptive measurement of the nature and location of prosodic segmentation information, using a reiteration paradigm, achieves several important aims: (i) it demonstrates the feasibility of such a paradigm - what is not a basic point, because of the metalinguistic task proposed to listeners; (ii) listeners are able to use the prosodic information carried by the reiterant speech since we have good association results of reiterant sentences with their originals, and discrimination of prosodic phrases that do not match the syntactic structure proposed. However, the task we proposed to the listeners is very complex and tests their ability to explicitly use their cognitive processing. We retrieve from such an experiment both cognitive information and indices about the dissociation clues they use.

We have already carried out some experiments with similar protocols as the dissociation experiment presented here, but using synthetic stimuli, reiterant, and original (for details, see [12]). We aim at studying the feasibility of such a paradigm to investigate the quality of a synthetic prosody and to extract a diagnostic analysis of the strengths and the weaknesses of a TTS system. An orthogonal experiment was also held by Morlec et al. [10], using stimuli with different prosodic patterns for one syntactic substrate. Listeners must dissociate the well and badly formed stimuli in a preference test. Instead of asking listeners to match directly the linguistics structures, they would separate them. These two paradigms perform the same objective by different ways, testing the adequacy of a prosodic continuum to describe a syntactic structure.

5. 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. Barbosa, P. & Bailly, G., "Characterization of rhythmic patterns for text-to-speech synthesis". *Speech Communication, Vol. 15*, p. 127-137, 1994.

3. Campbell, N., "Automatic detection of prosodic boundaries in speech". *Speech Communication, Vol. 13*, p. 343-354, 1993.

4. 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.

5. Hirshberg, J., "Intonation Theory", *Proceedings of the ESCA Workshop on intonation*, Athens, Greece, p. 19, 1997.

6. 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.

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

8. Lhote, E., Filleau, M. & Grange, F., "Reconnaissance de patrons intonatifs", *Proceedings 6ème Journées d'Étude de la Parole*, p. 29-37, 1975.

9. 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.

10. 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.

11. 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.

12. Rilliard, A. & Aubergé, V., "Reiterant Speech for the Evaluation of Natural vs. Synthetic Prosody", *Proceedings of the 3rd International Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998, (to appear).

13. 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.