Performance structures of mathematical formulae

Bleicke Holm (1), *Gérard Bailly* (1) *and Colette Laborde* (2)

(1) Institut de la Communication Parlée - CNRS/INPG/U3, (2) Laboratoire Leibniz - CNRS/INPG/UJF
46, avenue Félix Viallet, 38031 Grenoble Cedex 1, France

Abstract

We analyse here the prosodic structures of spoken algebraic equations read by one French speaker. The rhythmic structures of the utterances are characterised using the "performance structure" paradigm introduced by Grosjean et al. We show that the observed performance structures are qualitatively largely determined by syntactic structure and quantitatively largely influenced by phonotactic constraints especially for the attachment of mathematical operators. These data are then confronted to a superpositional model of intonation.

1. INTRODUCTION

To which extent a speaker can he convey hierarchical information by prosody? How are complex and deep syntactic structures encoded in the speech signal? These questions are essential to the comprehension of the language acquisition [14] and of particular interest in speech technology to help parsing and structuring the message. We study here the prosody of spoken mathematical formulae (MF) for several reasons: (1) they exhibit a deep and often ambiguous syntactic structure, (2) MF are complex structured drawn objects: dictation should carry information on the scope of operators in order that the listener may anticipate [10] the placement of operators and operands, (3) the enunciation of MF supposes a concept-to-speech generation process [11] where the speaker may use alternative prosodic and linguistic strategies for completing the task.

2. STATE OF THE ART

2.1. Mathematics as model for natural speech

At our knowledge very few works relate to the prosody of mathematics. This is certainly due to the fact that the syntax-to-speech mapping seems at the glance evident and that a direct application in speech technology is not of major economic interest. O'Malley et al. [16] try to recover by rules the parenthesis form of spoken formulae using the length of silent pauses. They state that there is no significant difference in the performances of mathematical experienced vs. naive speakers which indicates that the applied strategies are drawn from "natural" speech competence. The authors report that pauses tend to occur rather before the operators than after them. In our terminology (see section 4.3) operators tend to be attached at right. Using spoken mathematics as a model for "natural" speech reappears in the work of Streeter. A rather qualitative study on syntactically ambiguous sentences (Lehiste, Olive, Streeter [12]) is followed by a quantitative modelling of acoustic determinants used to disambiguate algebraic formulae [18].

2.2. Speech synthesis for mathematics

Raman [17] developed a complete document-to-speech synthesis system dealing with embedded mathematics. The system was designed as a speech browser of structured documents such as WEB pages for a blind user. He uses an original syntax-to-speech mapping involving a quasi-direct translation of tree structure into several so-called acoustic dimensions including prosodic dimensions but also more simple acoustic dimensions such as stereo.

The present study adopts a corpus-based approach where we aim at observing human performance and reaching the same level of intelligibility and comprehension by computers.

2.3. Performance structures

In line with several works dedicated to the relations between syntax and speech [4] Grosjean et al [8, 9, 13] propose an outstanding tool for representing the rhythmical structure. Performance structures organise break indices (BI) attached to word boundaries into a tree: the height of a node connecting two words (or group of words) corresponds to the value of the BI. Grosjean et al first imposed the speakers to pause between each word and used the pause duration as BI. As suggested by Duez [6] the duration of the last vowel of the preceding word was added to analyse normal speaking rates.

Our BI follows Barbosa et al proposals [2] that suggest that the rhythmical properties of French are governed by the occurrence of Perception Centres (PC) located in the vicinity of vocalic onsets. Successive PCs delimit an Inter Perceptual Centre Group (IPCG). The BI used in this study is a normalised lengthening factor of the IPCG duration $d(IPCG)$

containing the boundary. It is defined by

$$BI = NLF(IPCG) = \frac{d(IPCG) - \sum_{p \in IPCG} d_{av}(p)}{\Delta},$$

where $\Delta = d_{av}(vowels) + d_{av}(consonants) = 190ms$ accounts for an average IPCG duration. The average phoneme durations were computed from a separate speech database.

3. CORPUS

Our corpus was automatically generated using a formal description of equations [5] such as proposed in 4th grade mathematics exercises. They involve classical operations on 2nd degree polynomials.

The automatic generation of the corpus favours a compromise between different constraints: (1) the depth and syllabic weight of each node is systematically varied in order to ensure an optimal statistical coverage of the structural parameters supposed to influence the prosodic structure. This allows a systematic analysis of the utterances. (2) the ease of post-processing: we avoided too large groups of consonants, initial unvoiced stops and schwas. (3) the corpus size should be manageable : exponential explosion induced by the minimal pair principle implied by (1) is particularly true for the highly recursive syntax of MF.

We will not give here the formal description of the equation generator but rather the basic constructs. As in exercise books, each equation is introduced by "solve (equation)". Each equation is a function study $f(x)(=,>,<)g(x)$. Terminal nodes are polynomials of varying degrees (0,1 or 2). $g(x)$ is always a polynomial but the complexity of $f(x)$ is varied : a polynomial, products of polynomials and sums of products of polynomials. Prefixed operators (like $\sqrt{}$ or ||) can also operate on $f(x)$. Coefficients of the polynomials are integers and chosen randomly. In the same way we vary operators with identical syntactic function (like "+" and "−").

We end up with a corpus of 157 MF. The simplest and most complex MF are:

$$\begin{aligned} &solve: 6x < 1. \\ &solve: \sqrt{\tfrac{8x+9}{9x-9} - \tfrac{6x+9}{6x-6}} = 1. \end{aligned}$$

They were formatted by LaTeX and displayed on a computer screen. Our French speaker – a male PhD student – was instructed to enunciate the corpus without using lexical structural markers – such as "open parenthesis" – but instead making exclusive use of prosody. He read each MF until he was satisfied with two successive repetitions. Only the first was analysed here.

4. ANALYSIS

4.1. Global characterisation

The average speech rate is 3.8 syllables/s and the articulation rate – obtained by neglecting the silent intervals – is 4.6 syllables/s. The large difference between these two rates indicates that the speaker made use of many pauses and of a rather fast articulation rate which results in a strong rhythmical structuring of the utterances.

In the following we did not compute any BI for the last word of the utterance. In figure 1 we present the distribution of BIs for the entire corpus. The distribution spans from -1.8 to 7.8 with a maximum at -0.2.

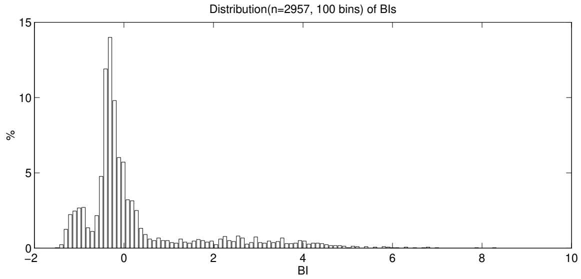

Figure 1: Distribution of BIs

How does the complexity of the MF affect the BI distribution? The four largest BI of each utterance are plotted as a function of the number of syntactic levels in figure 2. The largest BI becomes quickly stable. It is generally caused by the pause after "solve". On the other hand the distribution of other BIs depend more strongly on the MF complexity: it expands almost linearly as a function of the number of syntactic levels.

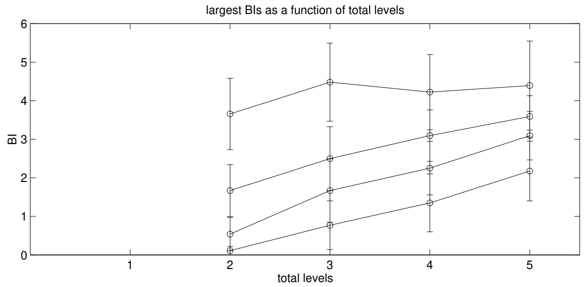

Figure 2: Means and standard deviations of the four largest BI versus total number of syntactic levels in the equations.

4.2. Performance structures

We present one typical example for the performance trees obtained from the data (see the upper tree of fig.3). When operators are merged with the its closest neighbouring word the organisation of operands follows quite nicely the syntactic hierarchy.

4.3. Attachment of operators

We will now focus on the behaviour of operators like "=" and the "+" in our example. From the tree we observe that the BI between "2)" and "=" is bigger than that between "=" and "2". We say: The operator "=" is attached at right. On the other hand the operator "+" is attached at left. As

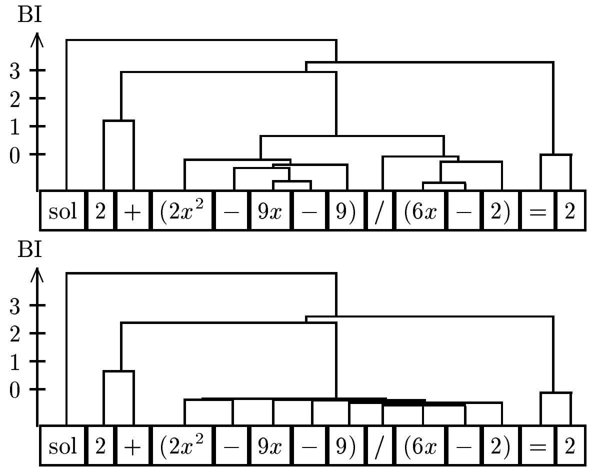

Figure 3: Performance trees (top: original, bottom: prediction)

a measure for the attachment of an operator (a op b), we define $A(op) = BI(op,b) - BI(a,op)$, i.e. $A(op) > 0$ means that "op" is attached at left.

In our corpus we observe a dependence of the attachment on the difference of number of syllables between left and right operands as shown in figure 4: the operator tend to be attached to the operand with smallest syllabic weight. Grammatical function words such as coordination and prepositions in natural language seem also to change their attachment as reported by Campbell [3]. In our data the syllabic lengths of the connected parts seem to play a major role.

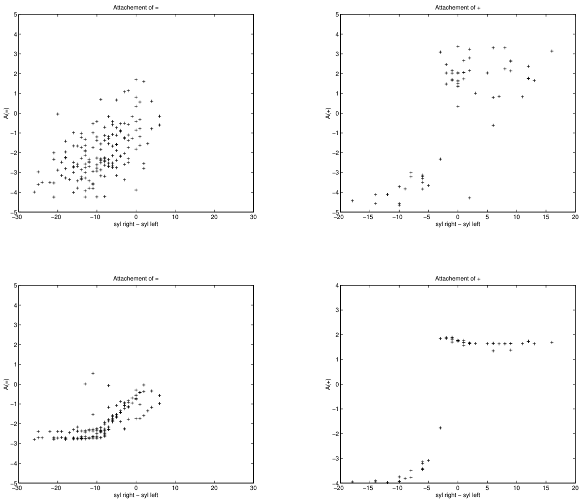

Figure 4: $A(op)$ as a function of the difference in number of syllables of the operands. Left: for "=", right for "+" operators (top: original, bottom: prediction (see 5.3)).

4.4. Discussion

The influence of number of syntactic levels on the BI distribution can interpreted in two ways: (a) BI depend on the length of the syntactic path from one word to the following as adopted by Cooper et al [4]: the tree height tend to grow as syntactic levels are introduced. (b) BI depend on the number of syllables of the delimited constituents : the number of syllables of constituents delimited by BI at a certain syntactic level tend also to grow as more syntactic sub-levels are introduced.

The hypothesis (a) supposes that the computation of each BI requires the knowledge of the whole syntactic tree. The hypothesis (b) supposes only the syllabic weights of the elements at each side of the juncture. Our data actually do not favour any of these hypothesis.

5. MODEL

The hypothesis (b) is in accordance with the ICP superpositional model of intonation where intonation and other linguistic agents share flexible layered synchronisation points (so-called "rendez-vous" [1]) between their constituents referred as landmarks in the following. We will test in the following if a model only based on this notion of layered synchronisation points can learn to reproduce the versatile attachment described above.

5.1. Description

We hypothesise that global prosodic patterns encode the organisation and function of constituents of each linguistic level (sentence, proposition, syntagms, words). We assume also independence: the patterns encoding a given level do not depend on the organisation of linguistic sub-levels. The prosodic contour is then obtained by a superposition of these prosodic patterns [7, 15].

These patterns are synchronised with the linguistic level by landmarks. These synchronisation landmarks are the boundaries between the constituents.

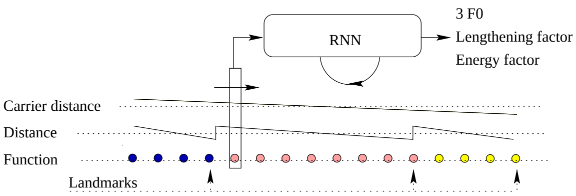

Figure 5: Flowchart of an RNN. For each IPCG of an utterance it delivers 5 prosodic parameters from the position and function of ongoing landmarks.

5.2. Implementation

This superpositional model of intonation is actually implemented using a linear combination of Recurrent Neural Networks (RNNs). Each RNN encodes the organisation of a particular linguistic level into prosodic trajectories (3 F0 values, a normalised lengthening factor and a energy factor per IPCG) from the specification of landmarks (see Fig. 5). For each IPCG of the utterance we specify (1) the function of the ongoing landmark (2) its absolute and relative position to the ongoing landmark in reference with the preceding one and also (3) its relative position to the ongoing landmark of the next higher "carrier" level if any. This information enables an anticipatory behaviour necessary for determining the importance of ongoing landmarks.

The levels considered in the study are: sentence-level (each formula consisting of two sentences: "solve" and the equation), group-level (left-hand side, operator, right-hand side), subgroup-level-1 (prefixed operator, operand) and subgroup-level-2 (high-level binary operator and its operands).

5.3. Predictions

Cooper [4], Zellner [19] and Grosjean [8] propose rule-governed models to predict the temporal structures; we use the hierarchical non-linear model described above for a multi-parametric generation of prosody.

The second tree of figure 3 shows the prediction for our example. Only the substructure of the product is not well predicted since our model does not yet treat this level.

We computed two kinds of prediction errors taking into account only the boundaries treated by the model. The first one furnishes a measure of the absolute error in BI prediction and is defined by:

$$E(BI) = \sqrt{\frac{\sum_{b\in B}(BI_d(b) - BI_p(b))^2}{\#B}},$$

where B: boundaries treated by the model, d: data, p: prediction and $\#B$: the number of elements in B. The second error measures how well the model reproduces the tree hierarchy. To each node we assigned a rank r: $r = 1$ for the largest BI, $r = 2$ for the second largest etc. Similarly we define $E(r)$. For the absolute error we obtain: $E_{mean}(BI) = 0.72$, $E_{median}(BI) = 0.62$, $E_{min}(BI) = 0.11$ and $E_{max}(BI) = 2.69$. For the hierarchical error the values are: $E_{mean}(r) = 0.26$, $E_{median}(r) = 0$, $E_{min}(r) = 0$ and $E_{max}(r) = 1.67$. The model reproduces exactly the hierarchy of 111 MF.

The figure 4 shows that our model fits suitable the data for the attachment of the operators.

6. CONCLUSIONS

Performance structure is a valuable tool for characterising the rhythmical structure of utterances that should be more heavily used in intonation studies. They evidence long-distance trading relations between constituents of the message. Such a rhythmical planning is of course of high importance in MF, where the syntactic and semantic predictability of constituents is very low.

Performance structures show that listeners may potentially recover fine details of the syntactic structures of MFs. Intelligibility tests should assess the effective use of this information. The next question will be how the listener actually process on line this information. We are currently experimenting a dictation task and analysing the phasing relations between the input speech and the output handwriting.

REFERENCES

1. Aubergé, V. La synthèse de la parole : "des règles aux lexiques". Thèse de troisième cycle, Université Pierre Mendès-France, Grenoble, France, 1991. Sous la direction de J. Rouault.
2. Barbosa, P. and Bailly, G. Characterisation of rhythmic patterns for text-to-speech synthesis. *Speech Communication*, 15:127–137, 1994.
3. Campbell, W.N. Automatic detection of prosodic boundaries in speech. *Speech Communication*, 13:343–354, 1993.
4. Cooper, W. and Paccia-Cooper, J. *Syntax and Speech*. Harvard University Press, Camebridge(MA), USA, 1980.
5. Drouhard, J.P. *Les Écritures symboliques de l'algèbre élémentaire*. PhD thesis, Université Denis Diderot - Paris 7, Paris, France, 1992. sous la direction de Daniel Lacombe.
6. Duez, D. *Contribution à l'étude de la structuration temporelle de la parole en français*. PhD thesis, Université de Provence, 1987.
7. Gårding, E. Intonation parameters in production and perception. In *Proceedings of the XIIth International Conference on Phonetic Sciences*, volume 1, pages 300–304, Aix-en-Provence, France, 1991.
8. Gee, J.P. and Grosjean, F. Performance structures: a psycholinguistic and linguistic appraisal. *Cognitive Psychology*, 15:411–458, 1983.
9. Grosjean, Grosjean, F., and Lane. The patterns of silence: Performance structures in sentence production. *Cognitive Psychology*, 11:58–81, 1979.
10. Grosjean, F. How long is the sentence? prediction and prosody in the on-line processing of language. *Linguistica*, 21:501–529, 1983.
11. Hayes, B. Speaking of mathematics. *American Scientist*, 84(2):110–113, 1996.
12. Lehiste, I., Olive, J., and Streeter, L. Role of duration in disambiguating syntactically ambiguos sentences. *Journal of the Acoustical Society of America*, 60:1199–1202, 1976.
13. Monnin, P. and Grosjean, F. Les structures de performance en français : caractérisation et prédiction. *L'Année Psychologique*, 93:9–30, 1993.
14. Morgan, J.L. and Demuth, K. *Signal to syntax: an overview*. Lawrence Erlbaum Associates, Mahwah, NJ - USA, 1996.
15. Morlec, Y., Bailly, G., and Aubergé, V. Generating prosodic attitudes in French: data, model and evaluation. In Botinis, A., editor, *Intonation: data, models and applications*, page (to appear). Cambridge University Press, Cambridge, UK, 1999.
16. O'Malley, M.H., Kloker, D.R., and Dara-Abrams, D. Recovering parentheses from spoken algebraic expressions. *IEEE Transactions on Audio and Electroacoustics*, 21(3):217–220, 1973.
17. Raman, T.V. *Audio Systems For Technical Readings*. PhD thesis, Cornell University, USA, 1994. sous la direction de David Gries.
18. Streeter, L. Acoustic determinants of phrase boundary perception. *Journal of the Acoustical Society of America*, 64(6):1582–1592, 1978.
19. Zellner, B. Pauses and the temporal structure of speech. In Keller, E., editor, *Fundamentals of speech synthesis and speech recognition*, pages 41–62. John Wiley and Sons, Chichester, 1994.