Keller, E., Zellner, B., Werner, S., & Blanchoud, N. (1993). The Prediction of Prosodic Timing: Rules for Final Syllable Lengthening in French. *Proceedings,* ESCA Workshop on Prosody (pp. 212-215). Lund, Sweden.

The Prediction of Prosodic Timing: Rules for Final Syllable Lengthening in French

Eric Keller, Brigitte Zellner, Stefan Werner, and Nicole Blanchoud
LAIP - Lettres, Université de Lausanne
CH-1015 Lausanne, Switzerland
Fax: +41 21 692 4639, Email: Eric.Keller@imm.unil.ch

ABSTRACT
Timing is an essential part of prosody, since it contributes to the semantic and syntactic modulations of speech conveyed by accent and intonation. Theoretical and empirical considerations suggest that temporal organization takes two main forms. First, it is a necessary corollary of accent placement and intonational modification. Second, it manifests itself in prolongations, pauses and hesitations related to sentence structure. A set of rules for the second set of temporal modulations is presented. In contrast to previous work, such rules depend only minimally on syntactic structures and can be formulated nearly entirely in simple phonological terms.

Rules for the prediction of the temporal structure of speech are important for the development of text-to-speech systems. Furthermore, such rules are of interest to our understanding of human linguistic functioning, particularly if they capture general principles of psycholinguistic operation. Previous work in this area has suggested that syntactic structures derived from psycholinguistic evidence (so-called "performance structures") could successfully predict the durations of *pauses* in speech (Gee & Grosjean, 1983). Specifically, pauses at major syntactic boundaries of this type tend to be longer than those at minor boundaries. Furthermore, *final syllable durations* adjoining such boundaries show similar correlations with syntactic boundary types. Psycholinguistic processing may thus employ "final syllable+adjoining pause duration" as a basic prosodic vehicle for marking certain hierarchical structures in speech.

The Monnin-Grosjean Rules

Several sets of rules of this type have been proposed for English, and an adaptation of these rules has recently been prepared for French (Monnin & Grosjean, in press). Fundamentally, the Monnin & Grosjean rules proceed as follows:

(1) *Nuclei* of prosodic constituents are identified from left to right: nouns, verbs and post-posed adjectives.

(2) *Prosodic constituents* are created by grouping words around the nucleus. Words are attached one by one. Special conditions determine whether words are attached to the right- or the left-lying nucleus.

(3) Word boundaries in the prosodic constituent are *indexed* to provide a measure of the *strength* or *importance* of boundaries between them. Basically, this is a count of the attachments (nodes) separating two words. However, various adjustments may intervene to handle special cases.

(4) *Higher prosodic constituents* are created to form a constituent hierarchy. This hierarchy is different from traditional syntactic hierarchies, since various "weight" parameters are taken into consideration, such as number of branching nodes.

(5) Higher prosodic constituents are indexed. The index count between two constituents is based on the number of nodes required to connect the constituents.

(6) *Further adjustments* may be required, depending on constituent and word length.

(7) Finally, by multiplication with a simple constant, index counts can be translated into *durations* of final syllable+pause segments.

A Verification of the Monnin-Grosjean Rules

As a first step, these rules were verified with respect to both the Monnin-Grosjean corpus and a new corpus. The verification of the original corpus permitted us to check our understanding

of the rules and to examine results in detail. The new corpus[1] represents three readings at different speech rates of three sentences by 12 speakers, 6 male, 6 female, 6 Parisian French, 6 méridional French (Toulouse) (7200 phonemes). The first reading was a practiced reading at normal speech rate, the second was a slow, deliberate reading, and the third was extra-deliberate (not used). The first sentence is syntactically and semantically complex, while the other two are quite simple. The corpus was manually labeled at the subphonemic level. Fo, energy and durations were measured at 10 ms-intervals.

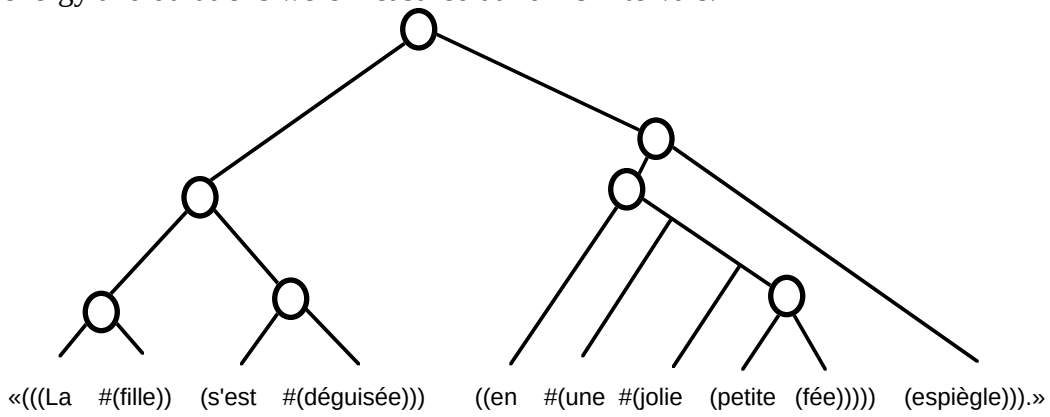

Figure 1. A performance structure tree according to the Monnin & Grosjean rules. A single node separates «petite» and «fée», which predicts a short final syllable and adjoining pause for «petite». By contrast, five nodes separate «déguisée» and «en», which predicts a much larger final syllable and pause duration at this major juncture. The #-mark specifies an attachment of grammatical to lexical words.

Results of the application to the Monnin-Grosjean rules were uneven (Table 1). While sentences 1 and 3 of the Caelen-Haumont data showed acceptable predictions in the .7 - .9 correlation range, sentence 2 showed particularly low predictions. With respect to Monnin and Grosjean's own data set, predictions for slow productions were found to be less successful than those for productions at a normal speech rate. This led to the following considerations:

(1) Monnin & Grosjean analyzed simple sentences that pose few problems of *hierarchical structure*. They suggest using general syntactic principles for *more complex structures.* (Our prediction for Caelen-Haumont's complex sentence 1 is in fact based on syntactic theory). However, reference to syntactic theory destroys the computational simplicity of the algorithm. Are complex structures really required?

(2) The application of the index count is *quite complex*. In particular, a number of minor adjustments are required to handle constituents of various lengths. Are all of these necessary? Are all of these adjustments "psycholinguistically real"?

The Keller-Zellner Rules

In view of the considerable existing literature and much experimentation performed on the two data sets, the following principles were formulated, and an algorithm was defined. The resulting rules satisfy our criteria of *simplicity*, respect of *psycholinguistic principles,* and *high predictive capacity* for the data sets at hand.

(1) *Prosodic constituents are formed on the basis of simple proximal syntax.* No syntactic structures more complex than those applying to a single phrase are required. This is considered to be "psycholinguistically simple" in the sense that children around 4-5 years show satisfactory prosodic grouping, an age at which they show an insufficient command of complex syntactic structures. Prosodic groups can be identified by the application of steps 1 and 2 of the Monnin-Grosjean rules.

(2) *Final syllable+pause durations increase in duration as the constituent proceeds.* The increase proceeds from an empirical minimum to an empirical maximum. The initial

[1] Kindly made available to us by Geneviève Caelen-Haumont, ICP/INPG, Université Stendhal, Grenoble, France.

hypothesis calls for equal steps. Increased durations correspond to a slowing down, which is a commonly observed phenomenon in speech.

(3) *Rhythmic alternance* was observed for two locations: post-verbally and in the middle of 4-6 word constituents. Rhythmic alternance occurs when one element is lengthened more than strictly required. As a consequence, the following element must be shortened "in order to conclude the constituent in time". Concretely, this amounts to postulating an *inversion of durations* for the word pair involved in the alternance.

The resulting algorithm is quite simple and is fully reproduced at the end of this paper. Correlations with the Caelen-Haumont and the Monnin-Grosjean data sets are reported in Table 1. It is found that correlations are quite regular. They never dip below a linear correlation of .7, and generally tend to be found in the .8 range.

An inspection of the evolution of Fo and energy values at the end of prosodic constituents postulated here shows some regularities. Fo values rise at the end of each constituent, except for the sentence-final constituent. Energy values fall regularly at the end of each constituent. This suggests that the temporal structure characterized here interacts directly with control over Fo and energy.

Table 1: Linear Correlations Between Predicted and Measured Final Syllable+Pause Durations According to Two Sets of Rules

	Monnin-Grosjean		**Keller-Zellner**	
Caelen-Haumont Data Set	**Normal**	**Slow**	**Normal**	**Slow**
Sentence 1	.786	.895	.862	.845
Sentence 2	.289	.375	.811	.829
Sentence 3	.925	.808	.878	.751
Mean	**.667**	**.693**	**.850**	**.808**
Monnin-Grosjean Data Set	**Normal**	**Slow**	**Normal**	**Slow**
Sentence 1	.890	.674	.873	.835
Sentence 2	.914	.796	.886	.954
Sentence 3	.981	.886	.773	.892
Sentence 4	.961	.826	.798	.850
Sentence 5	.947	.736	.827	.872
Sentence 6	.984	.711	.812	.835
Sentence 7	.931	.841	.754	.906
Sentence 8	.940	.585	.870	.809
Sentence 9	.968	.808	.701	.818
Mean	**.946**	**.763**	**.810**	**.863**

Conclusion

The performance of the new and simplified Keller-Zellner algorithm is encouraging. Proximal syntax can be used to create prosodic constituents, and final syllable+pause durations can be calculated using a simple set of rules. Text-to-speech systems could quite easily use these rules in conjunction with statistically determined values for non-final syllables (such as those proposed in O'Shaughnessy, 1984). It remains that the data considered here is limited. Only few sentences and speakers have been examined, and only read speech has been considered. Future research will automatize these rules and will examine predictions for larger and more varied data sets.

Acknowledgements

Grateful acknowledgement to G. Caelen-Haumont and F. Grosjean for making their corpora available and for offering their comments. Supported by the Office fédéral de l'éducation et de la science, Berne, for the ESPRIT project "Speech Maps" and the COST project 233 "Prosodics of synthetic speech".

The Keller-Zellner Algorithm

(1) Identification, from left to right, of the *nuclei* of the prosodic constituents: nouns, verbs and free-standing adjectives, adverbs and pronouns (such as "La chemise est *sale*", "c'est *bien*", "pense à *ça*").

(2) Creation of the *prosodic constituents* by grouping the words around the nucleus. All words to the left of the nucleus are attached to the right-lying nucleus, except for post-posed adjectives and post-posed pronouns which are attached to the left-lying nucleus ("la chemise *blanche*", "donne-*lui*").

(3) Calculation of *predictions for final syllable+pause durations*. Within each prosodic constituent, durations increase from a minimum to a maximum duration. Initially, the increase is assumed to occur in equal steps. (The minimum and maximum are assumed to be 50 and 350 ms in normal speech, 50 and 525 ms in slow speech.) The first final syllable in a constituent has a duration of minimum+step size ms.

(4) *Rhythmic tradeoffs:*

1. *Post-verbal trade-off:* When a constituent follows a verb and there are at least two words prior to the nucleus, the final syllable duration of the first word is lengthened with respect to that of the second word. (Exchange durations for words 1 and 2.)

2a. *Rhythmic alternance:* If a constituent is 4 or more words long, and if word 3 is 2 or more syllables long, word 2 is lengthened with respect to word 3. (Exchange durations for words 2 and 3.)

2b. *If rule 1 has already applied:* If a constituent is 4 or more words long, and if word 4 is 2 or more syllables long, word 3 is lengthened with respect to word 4. (Exchange durations for words 3 and 4.)

3. *Single-word constituents:* Constituents containing a single word show reduced final syllable durations. (Reduce durations for single word constituents by 50 ms.)

(5) *Measure of final syllable+pause.* The measure begins with the vowel of the final syllable and ends at the end of the pause. It includes whatever intervening consonant may occur, but it excludes the characteristic optional schwa of French méridional speakers (as in «biologiste**e**»). Excluding the optional schwa permitted us to make direct comparisons of data sets from northern and méridional speakers. Resulting time measures were very similar. For a limited data set, the intervening consonant was suppressed. However, resulting durations were found to show greater variability than those that included the consonant. Measures for sentence-final words were only known for a few sentences and were thus set to 0 in all cases for statistical purposes.

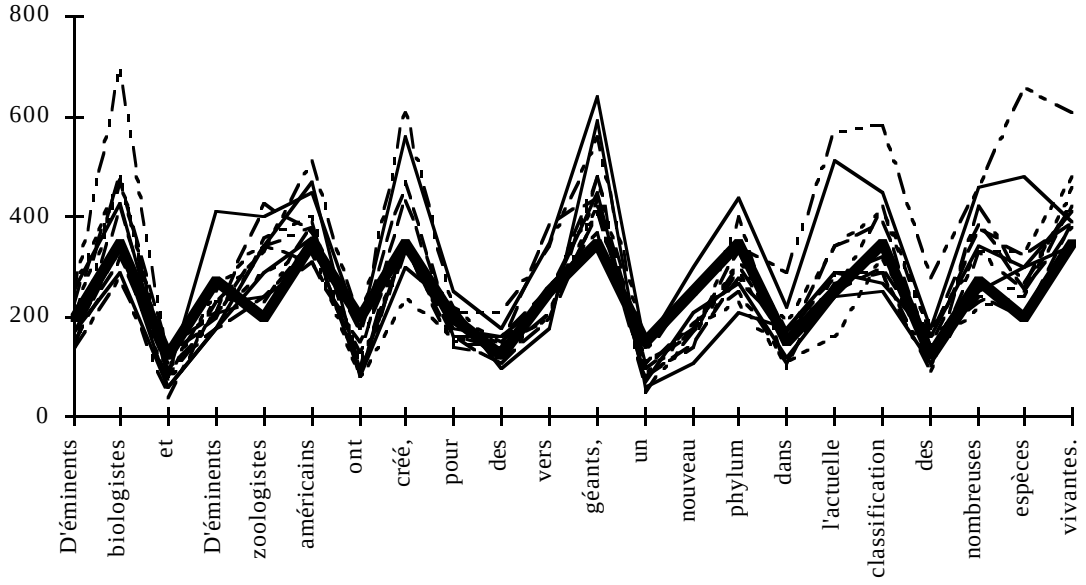

Figure 2. The prediction of the Keller-Zeller algorithm (thick line) for Caelen-Haumont's sentence 1 (thin lines: 12 speakers at normal speech rate).

References

Gee, J.-P., & Grosjean, F. (1981). Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology, 15,* 411-458.

Monnin, P., & Grosjean, F. (in press). Les structures de performance en français: caractérisation et prédiction. *L'Année Psychologique.*

O'Shaughnessy, D. (1984). A multispeaker analysis of durations in read French paragraphs. *Journal of the Acoustical Society of America, 76,* 6.