Zellner, B. (1998). Temporal structures for Fast and Slow Speech Rate. *ESCA/COCOSDA. Third International Workshop o* *Speech Synthesis*. (pp. 143 - 146), November 1998, Jenolan Caves (Australia).

TEMPORAL STRUCTURES FOR FAST AND SLOW SPEECH RATE

Brigitte Zellner

LAIP, IMM, Faculté des Lettres
University of Lausanne, 1015 Lausanne, Switzerland
Brigitte.Zellner.@imm.unil.ch

ABSTRACT

The rhythmic component in speech synthesis often remains rather rudimentary, despite recent major efforts in the modeling of prosodic models. The European COST Action 258 has identified this problem as one of the next challenges for speech synthesis. This paper is a contribution to a new, promising approach that was tested on a French temporal model.

INTRODUCTION

In a text-to-speech system, a well-constructed prosodic grammar is of primary importance. Extensive prosodic components have thus been incorporated into French speech synthesis systems (Bartkova, 1991; Keller, 1993; Sorin et al., 1987). However, within such efforts, relatively little work has concentrated on factors that control speech rhythm. The temporal dynamics of speech are often ill or insufficiently modelled in speech synthesis. This is a fundamental problem for speech synthesis and a major impediment to the improvement of the naturalness of synthetic speech.

1. THE MODELING OF SPEECH RHYTHM

Generally, the modeling of speech rhythm is reduced to the modeling of accentual structures. For example, temporal structures for French utterances are most often inferred from accentual structures, largely in the same manner as in English: the duration of each unit — syllable, GIPC or segment — is supposed to be directly related to its proximity to the accentual boundary (Barbosa, 1994; Beaugendre, 1995; Pasdeloup, 1992). In this view, temporal structures are presumed to be congruent with the accentual structure, regardless of the fact that French accent is different from English stress[1], and regardless of the fact that in this kind of approach, no other temporal events should intervene outside the accentual frame. Moreover, this standard approach is questionable since the distinction between stressed syllables and non-stressed syllables precedes the data prosodic analysis, instead of resulting from such a prosodic analysis.

In many languages, speech temporal structures are not exclusively shaped by the accentual event. In the proposed approach, it is suggested that other events are of importance for the modeling of temporal structures. For example in French, in a neutral reading task of declarative sentences, accentuation is mostly constrained by the phonosyntactic structures. This has generally been taken to be a sufficient basis for the prediction of durational events. But speech is temporally constrained by other fundamental components as well (for example bio-psychological, social and situational limits) (Zellner, 1997). It is claimed that a respect of these constraints — in addition to linguistic constraints — allows a better prediction of speech temporal structures. In fact, we have shown that the temporal structures for French neutral declaratives can be well predicted independently of the accentual structures, as long as they respect psycholinguistic principles (Zellner, 1996, 1998). The theoretical background of our approach is more largely described in author’s thesis and in this volume (Keller and Zellner), but it can be recalled that it is based on a vast number of psycholinguistic and phonetic studies, as well as on Levelt’s model (Levelt, 1989).

An illustration of this original approach is given with respect to variations of French speech rate. Changing the speech rate implies modifications at various levels of speech temporal structures (word groups, interlexical cohesion — “enchainements”[2], liaisons and pauses —, syllables and segments), which can be explained in terms of different temporal organisational strategies.

2. METHODOLOGY

2.1. Analysis

In our study, 50 declarative sentences were read by a highly fluent French speaker at two speech rates. As judged by ten native speakers of French, the readings were considered to be highly intelligible with no discernible dialectal accent. The signals were manually segmented and checked by two experts. Then, phonetic syllabic boundaries were controlled. In particular, the syllables containing a branching coda or a branching

[1] Accent for French and stress for English may have differing definitions (see for a complete review Caelen-Haumont, in press).

[2] Example of “enchainement”: “il est parti” (he’s gone). The /l/ of “il” is produced as the onset of the second syllable: /i lE parti/.

onset, and the cases of “liaisons” and “enchainements” effects were examined.

The raw syllabic durations were then normalised with a logarithmic transformation (Figure 1).

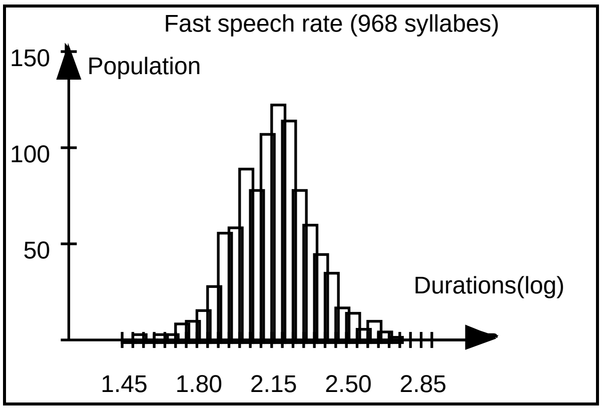

Figure 1. Syllabic durations after a log transformation (Fast speech rate)

Subsequently, for each speech rate, syllabic lengthenings and contractions were classified according to their deviations from the mean in this normalised space (Figure 2, at the end of this paper). The space is thus subdivided into five classes:

Class 1: Durations with major compression (more than 1 sd from mean duration)

Class 2: Durations with minor compression (between 1 sd and 0.5 sd from mean duration)

Class 3: Durations very close to mean duration

Class 4: Durations with minor lengthening (between 0.5 sd and 1 sd from mean duration)

Class 5: Durations with major lengthening (more than 1 sd from mean duration)

Generally, the higher a boundary is placed within the phonosyntactic or prosodic hierarchy, the more strongly it is marked for pitch and/or duration and/or energy. Since we assume that the temporal boundaries are marked by the syllable durations, the major groups should be marked by the greatest degree of contraction or lengthening — durations belonging to class 1 and class 5 — and the minor groups should be marked by relatively lesser deviations — durations belonging to class 2 and class 4.

2.2. Algorithm

Using this information, an algorithm for the prediction of temporal structures was developed (Zellner, 1996, 1997, 1998) that allows the Lausanne speech synthesiser LAIPTTS to read text in a fluent manner appropriate to either slow, normal or fast speech rate.

This relatively simple algorithm first looks for the distinction between lexical words (such as names, verbs, etc.) and grammatical words (prepositions, determinants, etc.). A minor boundary is applied each time a lexical word is followed by a grammatical word. A few supplementary rules are required when the number of lexical words in a prosodic group becomes too large, or when a series of other special conditions exists: fixed expressions, negationnal expressions, complex verbal expressions, etc. (Zellner, 1996). Then a major boundary is applied each time a punctuation mark is found. Another major boundary is placed in the middle of the longer sentences, on the closest minor boundary.

Depending on the speech rate, a number of modifications are then applied to this temporal segmentation. According to the context, five rules are useful for fast speech (for more details, see Zellner, 1998):

a. Creation of “enchainements” (interlexical links)
b. Elision of schwas
c. Addition of schwas (rhythmic equilibrium)
d. Addition of pauses (long sentences)
e. Addition of temporal boundaries (long sentences)

For slow speech, five rules are applied (for more details, see Zellner, 1998):

a. Addition of syllables (final schwas in words)
b. Dieresis
c. Creation of “enchainements”
d. Addition of pauses
e. Addition of groups boundaries

This algorithm, coded in C, was used to transform the phonetic chain into the phonological structure required for fast or slow speech rate. 50 automatically segmented sentences at the two speech rates were then compared to those produced by the natural speaker, according to the measured syllabic durations (see figure 2).

3. RESULTS

The results are very promising, since the different "phonological" structures (which means in that case: syllabic chain, temporal boundaries, pauses) could be predicted correctly for the two speech rates, without distinguishing stressed and unstressed syllables.

Moreover, based on this segmentation, a durational model calculated with a general linear model gave very promising results (similar to the Keller-Zellner, 1995, 1996). The correlations between predicted and measured units for the duration of syllables were at least 0.72 (p<0.001) and for segments 0.74 (p<0.001) at the two speech rates.

Fast Speech Rate	Predicted	Measured
Enchaînements	47	44
Pauses	16	14
Syllables	964	968

Schwas	106	114
Minor boundaries (+ pause)	122 (9)	128 (8)
Major boundaries (+ pause)	34 (7)	35 (6)
Total number of boundaries	156	163

Slow Speech Rate	Predicted	Measured
Enchaînements	31	34
Pauses	48	50
Dieresis	4	5
Syllables	997	1001
Schwas	135	120
Minor boundaries (+ pause)	120 (40)	121 (41)
Major boundaries (+ pause)	47 (8)	43 (9)
Total number of boundaries	167	164

4. DISCUSSION

The prediction of temporal structures is made on the basis of concepts which respect most of the temporal constraints of speech and which feed directly into a simple algorithm of word grouping. The underlying model is relatively simple and generalises easily to other speakers. For declarative sentences, temporal structures can be identified in any database by means of a simple statistical analysis.

A major benefit of this work is to clarify the modeling of temporal structures in French TTS systems. It is demonstrated that for French declarative sentences, temporal structures may be well predicted without looking for stress / accent events.

I would not claim that accentuation has nothing to do with timing. I suggest that in certain situations such as in neutral speaking, *major* temporal events are marked elsewhere in the speech chain than in places of accentuation. This is an important claim if one wants to improve the naturalness of speech synthesis, in particular the fluency of synthetic speech.

Further improvements of naturalness in speech synthesis are conceivable, once these temporal structures are integrated, together with accentuation and melodic structures, into a complete prosodic model.

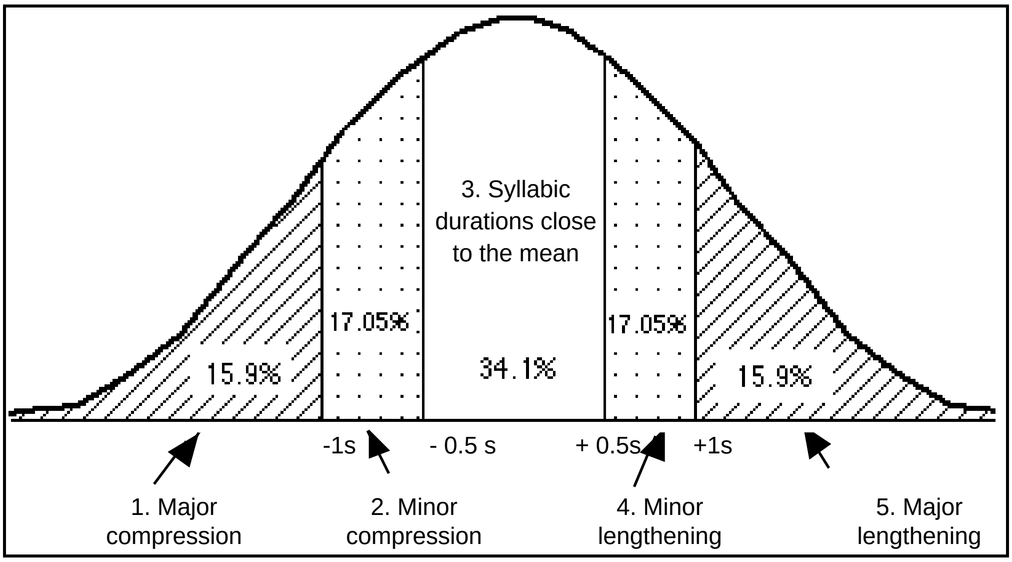

Figure 2. Subdivision of syllabic durations after normalisation.

REFERENCES

Barbosa, P. A. (1994). *Caractérisation et génération automatique de la structuration rythmique du français*. Thèse de Doctorat. U.R.A. CNRS n°368 - INPG/ENSERG, Université Stendhal, Grenoble.

Bartkova, K. (1991). Speaking rate in French application to speech synthesis. *XIIème Congrès International des Sciences Phonétiques*, (pp. 482-485). Aix en Provence. Actes.

Beaugendre, F. (1995). Generating French intonation at different speaking rates. *ESCA. Eurospeech'95*. European Conference on Speech Communication and Technology. Madrid, September. ISSN 1018-4074.

Caelen-Haumont, G. (to appear). Prosodie et sens, une approche experimentale. Ed. du CNRS.

Keller, E., Zellner, B., Werner, S., and Blanchoud, N. (1993). The prediction of prosodic timing: Rules for final syllable lenthening in French. *Proceedings ESCA Workshop on Prosody, September 27-29. Lund, Sweden.* 212-215.

Keller, E., & Zellner, B. (1995). A statistical timing model for French. *XIIIème Congrès International des Sciences Phonétiques, 3* (pp. 302-305). Stockholm.

Keller, E., & Zellner, B. (1996). A timing model for fast French. *York Papers in Linguistics*, 17, University of York. 53-75.

Keller, E. & Zellner, B. (1998). Motivations for the Prosodic Predictive Chain. In this volume.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MIT Press. Cambridge.

Pasdeloup, V. (1992). Durée intersyllabique dans le groupe accentuel en français. *Actes des 19émes Journées d'Etudes sur la Parole*. (pp. 531-536). Bruxelles.

Sorin, C., Larreur, D. & Llorca, R. (1987). A rhythm-based prosodic parser for text-to-speech systems in French. *XIème Congrès International des Sciences Phonétiques* (pp. 125-128). Talinn, Estonia. Proceedings.

Zellner, B. (1996). Structures temporelles et structures prosodiques en français lu. *Revue Française de Linguistique Appliquée: La communication parlée.* 1. (pp.7-23). Paris.

Zellner, B. (1997). Fluidité en synthèse de la parole. In E. Keller, & B. Zellner (Eds.), *Les défis actuels en synthèse de la parole*e, *Études des Lettres, 3.* (pp. 47-78). Université de Lausanne.

Zellner, B. (1998). *Caractérisation et prédiction du débit de parole en français. Une étude de cas.* Thèse de Doctorat. Faculté des Lettres, Université de Lausanne.

ACKNOWLEDGEMENTS

With many thanks to Eric Keller (LAIP) for his welcome help, and the precious remarks of my colleagues of COST 233 and COST 258. Portions of this research have been supported by the Office Fédéral pour l'Enseignement et la Recherche (OFES) through its support to our research activities under COST 233 and COST 258.