Revisiting the Status of Speech Rhythm

Brigitte Zellner Keller

Laboratoire d'Analyse Informatique de la Parole
Faculté des Lettres, UNIL,
CH-1015 Lausanne, Switzerland
Brigitte.ZellnerKeller@imm.unil.ch
Online Synthesis: http://www.unil.ch/servlets/imm/SyntheseServlet
Sound examples: http://www.unil.ch/imm/docs/LAIP/LAIPTTS.html

Abstract

Text-to-Speech synthesis offers an interesting manner of synthesising various knowledge components related to speech production. To a certain extent, it provides a new way of testing the coherence of our understanding of speech production in a highly systematic manner. For example, speech rhythm and temporal organisation of speech have to be well-captured in order to mimic a speaker correctly.

The simulation approach used in our laboratory for two languages supports our original hypothesis of multidimensionality and non-linearity in the production of speech rhythm ([19], [20], [21]. This paper presents an overview of our approach towards this issue, as it has been developed over the last years.

We conceive the production of speech rhythm as a multidimensional task, and the temporal organisation of speech as a key component of this task (*i.e.,* the establishment of temporal boundaries and durations). As a result of this multidimensionality, text-to-speech systems have to accommodate a number of systematic transformations and computations at various levels. Our model of the temporal organisation of read speech in *French* and *German* emerges from a combination of quantitative and qualitative parameters, organised according to psycholinguistic and linguistic structures. (An ideal speech synthesiser would also take into account subphonemic as well as pragmatic parameters. However such systems are not yet available).

Introduction

On the basis of empirical studies for read speech, Zellner Keller and Keller have conceived a model of speech rhythm for speech synthesis. Our originality resides in two points:

• The temporal organisation of speech plays a major role in the recreation of rhythm ([7], [8], [9], [10], [17], [18], [20], [22], [24]). Any disorganisation of temporal structures rapidly leads to impairments of speech rhythm. For example, temporal boundaries occurring at "the wrong places", or odd durations of segments in "certain places" of the verbal stream, may disrupt tremendously speech fluency. Assuming that durations are to a certain extent associated with the temporal structures, and in order to establish what the significant parameters are for the prediction of these structures, we used a "pure" statistical approach towards the temporal organisation of speech, *i.e.,* independent of any linguistic categories other than the concept of phones and phonetic syllables. Parameters incorporated into our model are thus statistically well-motivated.

• Rhythm is a multidimensional and non-linear cognitive construction ([19], [20], [10], [25]). These properties are particularly important at the temporal level and in our model, temporal organisation of speech is *not only a question of phone durations.* Other systematic constraints have to be incorporated as well.

The overview of our temporal model begins by considering the word grouping issue, and it will pursue by examining the multi-layer modelling of durations

The Word Grouping Issue

This work is fully described in [9], [20], [21]. The simulation of speech rhythm requires first of all a realistic imitation of the manner in which speakers organise their information flow in time, *i.e.*, the word grouping strategy. Word groups are the building blocks for the reconstruction of the temporal organisation of speech. If this word grouping strategy is not correctly handled, temporal boundaries are misconceived, which renders spoken utterances odd or even disfluent. Our basic assumption is that the process controlling spoken information flow is partly related to the temporal dimension. Assuming that the temporal structure of speech is given by durations of speech units, the issue becomes understanding the relationship between such durations and the overall temporal structure.

1.1. Looking at the durations[1]

First, it is worthwhile to emphasize the efforts expended in our laboratory to obtain robust data and to perform a *careful manual segmentation* of speech units, on the basis of an explicit protocol (see annex 4, [20]). On a representative corpus obtained in the mid-1990s, interjudmental agreement was scored for boundary identification between 1 (low agreement) to 3 (excellent agreement), and for accuracy between 1 (more than 2 F0 periods difference) to 3 (less than 1 F0 period difference). Over 50 types of segmental transitions, there were no cases of score 1. The average agreement was 2.53 for boundary identification and 2.68 for accuracy [8].

The second step is to look at the statistical distribution of the segmental and syllabic durations. Raw syllabic and segmental durations tend to show distributions that deviate from a Gaussian curve. After various explorations, the logarithmic transformation was found to be an excellent candidate for this normalisation in our data, as shown for example in figure 1.

[1] Various empirical studies of Keller and Zellner Keller on different corpora are discussed in this paper. For details on methodology and results, please refer to the relevant publications.

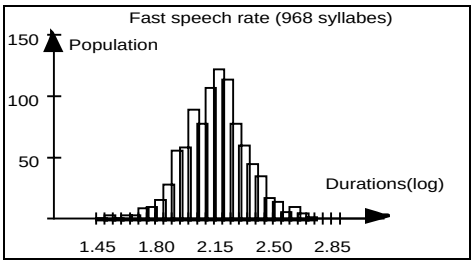

Figure 1. Syllabic durations after a logarithmic transformation

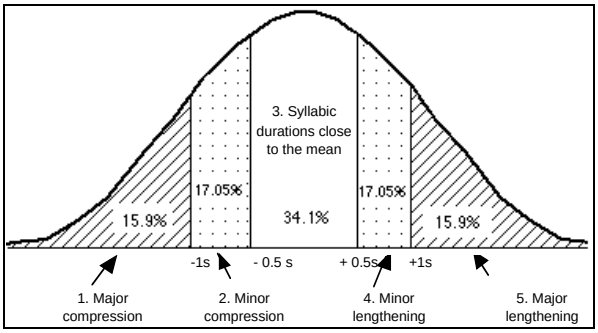

Figure 2. The theoretical distribution of syllabic durations after logarithmic normalisation. This distribution is subdivided into five syllabic duration classes.

After this normalisation, syllabic durations are classified in relation to the mean duration (Figure 2). The benefit of this classification is that *no linguistic a-priori concept is introduced in the course of the analysis* (such as stressed / non stressed syllables). This provides a better chance of identifying parameters that are statistically significant for the prediction of the temporal organisation of speech.

Table 1: Samples of normalised syllable durations distributed over five classes for two speech rates.

5 classes of syllabic durations	*Fast speech rate* (n = 968)	*Slow speech rate* (n = 1001)	Theoretical normal distribution
Class 1 More than -1 σ (very much shortened)	14.0%	14.7%	15.9%
Class 2 between –1 σ and -0.5 σ (shortened)	15.3%	13.3%	17.05%
Class 3 close to the mean	41.3%	40.5%	34.1%
Class 4 between 0.5 σ and 1 σ (lengthened)	13.6%	15.8%	17.05%
Class 5 More than 1 σ (very much lengthened)	15.8%	15.7%	15.9%

A study of 50 sentences produced at two speech rates by the same speaker showed that this statistical classification captures syllable durations adequately (see table 1), all while admitting that both rates show some non-significant "bunching" close to the mean ($p < 0.0001$).

We hypothesised that syllables of certain duration class are associated with major groups, and others with minor groups [19]. Minor temporal boundaries would be signaled by syllable durations with a minor deviation from the mean (class 2 and 4), and syllable durations close to the mean (class 3) would characterise most of the non-temporal boundary syllables (see figure 2 and 3). Since a temporal major group is bounded by two major boundaries, we theorised that syllable durations with a major deviation (class 1 and 5) would frame a major group, which in turn may contain one or several minor groups.

We found this prediction to be so strong for French that a first hypothesis for minor and major temporal boundary identification could be based on this criterion [20]. We also found that temporal boundary placement interacts with lexical elements in French..There was evidence of regular relations ($r > 0.90$ for fast and slow speech rates) between.:

• Types of words (lexical, grammatical) and temporal boundaries (*i.e.*, type of durations). For example, grammatical words tend to start a group; lexical words tend to end a group

• Pauses and temporal boundaries (*i.e.*, type of durations).

In other words, the information flow along the time line emerges from:

• *minor building blocks*: grammatical and lexical words, which are combined according to phonosyntactic rules;

• *major building blocks*: a tendency to equilibrate the length of phrases and to insert a pause around the middle of an utterance (Figure 4).

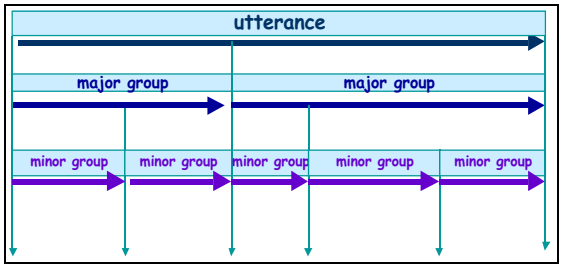

Figure 3. The hypothetic temporal organisation of an utterance

In that sense, all our durational studies ([7], [8], [18], [19], [20], [21]) confirm the results of previous psycholinguistic studies of word grouping ([1], [3], [4], [5], [6], [17]).

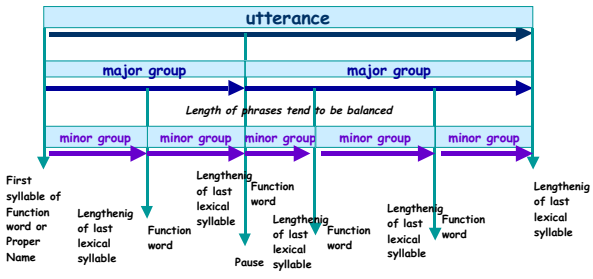

Figure 4. Relationship between the temporal organisation of an utterance and word grouping.

1.2. Prediction of boundaries and durational classes

Based on this analysis, an algorithm for the prediction of temporal boundaries was constructed ([7], [19]). This simple model predicts the durational class for each syllable (Table 1). For example, if a syllable is predicted to be of class 4, its duration is expected to be lengthened within the limited range of 0.5 σ to 1 σ.

The 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, negational expressions, complex verbal expressions, etc. [19]. 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 ([20], [21]).

The outcome of this algorithm (Figure 5) gives a "temporal skeleton" of the utterance, *i.e.*, a coarse profile of the syllabic durations ([23]).

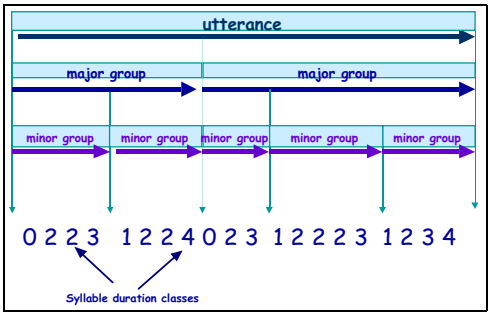

Figure 5. Outcome of the word grouping prediction: a temporal skeleton

Prediction of speech unit durations

This work is fully described in ([8], [9], [10],[20]). Once the temporal skeleton is predicted, durations have to be predicted. We used a step-wise approach, predicting durations with a statistical model and then adjusting these durations to phonological and serial constraints.

2.1 Statistical model

Our initial durational model was conceived on the basis of the analysis of 100 read sentences on the one hand, and 50 read sentences on the other hand, produced at two speech rates by the same speaker for French.

A modified step-wise statistical regression technique for segmental, syllabic and phrase level information was used to develop this model of the speaker's timing behaviour. First, at the segmental level, segmental durations were grouped in 6 to 8 groups on the basis of their mean and std. deviation in log ms. Predictions made on those groups were more robust. Then a general linear model (GLM) was applied for the prediction of segmental durations ("intrinsic duration"). The residue (predicted - measured duration) served as the basis for further factorial modelling at the syllabic and phrasal level (for an example, see Table 2).

The significant parameters incorporated at the segmental level were the durational group of the current phone and adjacent segments. At the syllabic level, parameters were: type of word, type of syllable, position in the temporal group, and phonotactic constraints. At the phrasing level, the position in the phrase was taken into account.

Table 2. GLM applied to the prediction of syllable durations at the fast speech rate

• Prediction of durations (Zellner, 1998)
Example for fast speech rate

Source	df	SumsSquares	MeanSquare		F-rapport Prob
X_0	1	4559.51	4559.51	454865	≤ 0.0001
X_1	157	21.4501	0.136625	13.63	≤ 0.0001
X_2	2	3.1511	1.57555	157.18	≤ 0.0001
X_3	4	1.4585	0.364624	36.376	≤ 0.0001
X_4	1	0.441099	0.441099	44.005	≤ 0.0001
X_5	1	0.493451	0.493451	49.228	≤ 0.0001
X_6	1	0.074769	0.074769	7.4591	0.0064
error	815	8.16946	0.010024		
total	981	35.2385			

Syllabic duration (log10(ms)) = constant
+ coefficient type of segment * type of syllable
+ coefficient type of temporal boundary
+ coefficient number of segments in the syllable
+ coefficient presence or absence of schwa in the syllable
+ coefficient grammatical / lexical word
+ coefficient length of the word

At this stage of the model, correlations between measured durations and predicted durations were r=0.876 (n=981; p<0.001; RMSE = 18 ms) for fast speech rate and r=0.83 (n=1049; p<0.001; RMSE = 22.9 ms) for slow speech rate. However, the prediction of durations needs to be completed

2.2. Phonological adjustment

This work is fully described in [20]. The predicted syllabic duration must then be distributed among the phones according to a trade-off between the target duration: i.e. the predicted syllabic duration S' and the sum of "intrinsic" segmental duration S (durational group of segments, related to the speech rate: see section 2a). To solve this trade-off (S' - S), we examined the relationship between the durational class of syllables (class 1 to class 5) and phonological syllable structures.

Pearson correlation coefficients show that the duration of each part of a syllable (onset "O", nucleus "N", coda "C") is more or less correlated to the class of syllabic duration in which it occurs. The trade-off (S' - S) is thus solved according to the phonological syllabic type, ordering the hierarchy into the distribution of durations:

• in syllables ON, NC: the first part is highly correlated to the syllabic durational type;

• in syllables ONC: the three parts are equally mildly correlated to the syllabic durational type

In summary, durations are predicted on the basis of a general linear model adapted to the syllabic structure. At this stage, durations are still predicted without taking account of durational interactions with adjacent syllables: the serial constraint, which we turn to last.

2.3. Serial constraint

Durational serial interactions are not generally incorporated into contemporary predictive models of timing for speech synthesis, at least with respect to data-based models used for speech synthesis (e.g., [2]: artificial neural network; [14]: classification and regression tree; [12]: multiplicative and compositional model; [16]: sum-of-products; [8]: general

linear model). These sophisticated statistical and artificial neural network models typically draw their predictive information from phonological, syntactic and semantic, but not from durational parameters. However, since speech is a rhythmical activity, it can be expected that certain serial interactions act as a kind of span, structuring the speech flow.

In a recent study [10], we identified a durational anti-correlation component that manifested itself reliably within 500 ms, or at a distance of one, and possibly a second syllable. This effect means that within a 500 ms window, if a syllable tend to be short, the following syllable will tend to be longer. For example, we found a minor anticorrelation of r=0.234 for fast speech rate. Although numerically minor, a computer implementation of this component resulted in noticeable perceptual effects.

Conclusion

Simulating the production of speech rhythm reveals the complexity of this task as well as the dependence on an extensive set of predictive quantitative *and* qualitative parameters.

The originality of our computational simulation is grounded at two levels. First, the temporal component plays a major role in our reconstruction of speech rhythm, and provides the temporal skeleton. Second, the outcome of temporal modelling results from the harmonisation of various layers (segmental, syllabic, phrasal). This permits to impose differentiated effects at various levels, which is the key to speech coherence and speech fluency.

Readers are invited to test the online speech synthesis of the LAIP (http://www.unil.ch/servlets/imm/SyntheseServlet) and to judge by themselves how our original approach produces a realistic reconstruction of speech rhythm for read speech. Moreover, adaptation of these principles and algorithms to German show that this approach is also applicable to other languages [15].

Even though temporal issues are not at the center of current interest in Prosody, it is hoped that this contribution has illustrated the scientific interest of these matters.

Acknowledgements

Grateful acknowledgement is made to the Office Fédéral de l'Education for supporting this research through its funding of Swiss participation in COST 258, to the État de Vaud and to the University of Lausanne.

References

[1] Boomer, D.S., & Ditmann, A.T. Hesitation pauses and juncture pauses in speech. *Language and Speech*, *5*, 215. (1962).

[2] Campbell, W.N. (1992). Syllable-based segmental duration. In G. Bailly, & al (Eds.), *Talking Machines. Theories, Models, and Designs* (pp. 211-224). Elsevier Science Publishers.

[3] Chafe, W. (1980). Some reasons for hesitating. In W. Dechert & M. Raupach, (Eds.), *Temporal variables in speech* (pp. 169-180). Mouton.

[4] Cook, M., Smith, J., & Lalljee, M. (1974). Filled pauses and syntactic complexity. *Language and Speech, 17*, 11-16.

[5] Goldman–Eisler, F. (1972). Pauses, clauses, sentences. *Language and Speech*, *15*, 103-113.

[6] Grosjean, F., & Deschamps, A. (1975). Analyse contrastive des variables temporelles de l'anglais et du français. *Phonetica, 31*, 144-184.

[7] 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.

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

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

[10] Keller E., Zellner Keller, B. & Local, J. (2000). A serial prediction component for speech timing. In Sendlmeir, W. (Ed). Speech and Signals. Aspects of Speech Synthesis and Automatic Speech Recognition. (pp. 41 - 49). *Forum Phoneticum, 69.* Frankfurt am Main: Hector.

[11] Keller, E., & Zellner-Keller, B. (to appear). Speech Synthesis in Language Learning: Challenges and Opportunities. Proceedings of InSTIL 2000 Conference. Dundee, Scotland. (to be reprinted as a volume).

[12] Ogden, R. Local, J. & Carter, P. (1999). Temporal interpretation in ProSynth, a prosodic speech synthesis system. *Proceedings of the XIVth International Congress of Phonetic Sciences* (Ohala, J.J., Hasegawa, Y., Ohala, M., Granville, D., and Bailey, A.C. eds.), 2, pp. 1059-1062. University of California, Berkeley, CA.

[13] Riedi, M. (1998). *Controlling Segmental duration in Speech Synthesis Systems*. Ph.D. Thesis. ETH. Zürich. (TIK-Schriftenreihe 26). 174 pp.

[14] Riley, M. (1992). Tree-based modelling of segmental durations. In G. Bailly et al., (Ed.). *Talking Machines: Theories, Models, and Designs* (pp. 265 - 273). Elsevier Science Publishers.

[15] Siebenhaar, B., Zellner Keller, B., Keller, E. (2001). Phonetic and Timing Considerations in a Swiss High German TTS System. in Keller, E., Bailly, G., Monaghan, A., Terken, J. and Huckvale, M. Eds. Improvements in Speech Synthesis. Chichester: John Wiley.

[16] van Santen, J.P.H (1993). Timing in text-to-speech systems. *Proceedings of the 3rd European conference on speech communication and technology* (pp. 1397-1404). Berlin.

[17] Zellner, B. (1992). Le bé bégayage et euh... l'hésitation en français spontané. *Actes des 19ème Journées d'Etudes sur la Parole (J.E.P*), Bruxelles. 481-487.

[18] Zellner, B. (1994). Pauses and the temporal structure of speech, in E. Keller (Ed.) Fundamentals of speech synthesis and speech recognition. (pp. 41-62). Chichester: John Wiley.

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

[20] Zellner, B. (1998.a). Caractérisation et Prédiction du Débit de Parole en Français. Une étude de cas. Thèse de Doctorat. Faculté des Lettres, Université de Lausanne.

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

[22] Zellner Keller, B. (2001). La modélisation du rythme de parole. Le problème de la cohérence temporelle.Oralité et Gestualité. Interactions et comportements multimodaux dans la communication. Actes.(pp. 640 - 645). Aix en Provence (France).

[23] Zellner Keller B. (2001). Ecrire le Rythme de la Parole. *Parole.* 17-18-19, Aix (France).

[24] Zellner Keller, B., Keller, E. (2001). A non-linear rhythmic component in various styles of speech. in Keller, E., Bailly, G., Monaghan, A., Terken, J. and Huckvale, M. Eds. Improvements in Speech Synthesis. (pp. 284-291). Chichester: John Wiley.

[25] Zellner Keller, B., & Keller, E. (to appear). The chaotic nature of speech rhythm. In Ph. Delcloque and V. M. Holland (Eds). Integrating Speech Technology in Language Learning. Swets & Zeitlinger.