A Serial Prediction Component for Speech Timing

Eric Keller (1), Brigitte Zellner-Keller (1), & John Local (2)

(1) Laboratoire d'analyse informatique de la parole (LAIP), IMM-Lettres, University of Lausanne, 1015 Lausanne, Switzerland
Email Eric.Keller@imm.unil.ch, Brigitte.ZellnerKeller@imm.unil.ch,
http://www.unil.ch/imm/docs/LAIP/LAIP.html
(2) Department of Language and Linguistic Science, University of York, Heslington, York YO10 5DD, United Kingdom
Email lang4@york.ac.uk, http://www.york.ac.uk/~lang4

Abstract

Durational serial interactions are not generally incorporated into contemporary predictive models of timing for speech synthesis. In this study, an anti-correlational factor applied at the syllable level was identified for syllable lags occurring within roughly 500 ms. As applied to synthetic speech, a strongly anti-correlational effect appears to lend a pleasant, "swingy" effect to the speech output, while the absence of such an effect results in a more "regimented" style of speech. The first may be appropriate for a standard reading of continuous text, and the latter more appropriate for other styles of speech, such as train announcements.

1. Introduction

During the second half of the 20th century, the serial dimension of speech has been incorporated into a large number of phonetic descriptions. For example, coarticulation effects have been widely described, and HMM models exploit serial information extensively for speech recognition purposes.

Somewhat curiously, durational interactions along the serial dimension have not yet been integrated into predictive models of the temporal organisation of speech, at least with respect to data-based models used for speech synthesis (*e.g.*, Campbell, 1992: artificial neural network; Riley, 1992: classification and regression tree; Ogden *et al.*, 1999 multiplicative and compositional model; Van Santen, 1992: sum-of-products; Keller & Zellner, 1995: general linear model). By far most current temporal models used in contemporary speech synthesis systems limit serial interactions to "time-less" variables. 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 temporal harmonic attractors appear along the speech stream (see Zellner-Keller & Keller, to appear). A temporal serial interaction component may thus be considered as a kind of *span*, structuring the speech flow.

In this paper, the temporal serial dependency in speech temporal organisation is assessed for French, and to a lesser extent for English. A weak, but statistically significant serial position effect is documented for both languages. Theoretical issues arising from these findings are discussed with respect to improvements in the naturalness on speech synthesis.

2. Description of the Serial Dimension

The essential question addressed here concerns the capacity of durations $x_1...x_k$ to predict durations x_{k+1} in a chain $x_1...x_k|x_{k+1}$, where the members of the chain are relevant units of time (typically a syllable or segment). Ideally, one would wish the chain to be of considerable length, in order to be able to assess the effect of lag, *i.e.*, the question of the number of elements over which the potential influence makes itself felt.

In the present study, the *syllable* was chosen as unit of time. This is in accordance with the common observation that at least at one grain of analysis, speech timing variations typically extend over the width of a syllable. In a language like English, for example, all segments constituting a syllable tend to be shortened to some degree in unstressed, and tend to be lengthened to some degree in stressed conditions.

There are also several disadvantages to choosing the syllable as a unit of time. First, few utterances contain enough syllables for a statistically satisfactory evaluation of the question of *lag*. Second, the theoretical definition of a syllable is not perfectly clear or predictable, and the question can not always be settled by an examination of the actual speech output. To take examples from English, is the syllabification "in-fir-mi-ty" or is it "in-firm-i-ty"? Or is the word syllabified "fe-ver-ish" or "fev-rish", *i.e.*, does it contain three or two syllables? And the final problem is that there are few precisely (*i.e.*, manually) segmented speech databases that list syllable boundaries.

As it happens, databases developed in our two laboratories satisfied these criteria to a considerable extent. Segmentation had been performed manually with considerable care and reverification, syllable boundaries had been specified, and questionable cases had been settled by an auditory and visual examination of the given signal portion (for more details, see Zellner, 1998a,b; Ogden et al., to appear). Also, with respect to the French database, utterances were sufficiently long to permit interesting statistical analyses of the question of lag.

3. Serial Dependency

Essentially all current timing models for speech synthesis incorporate at least one type of serial dependency, the dependency on the *identity* of the preceding and succeeding sound. This reflects well-known phonetic interactions between adjoining sounds, such as the fact that in many languages, vowels preceding voiced consonants tend to be somewhat longer than similar vowels preceding unvoiced consonants ("bead" *vs.* "beat"). A number of other commonly used predictive parameters (*e.g.*, lengthen syllable duration at the end of phrases and sentences) can also be reinterpreted as serial dependencies (*i.e.*, "lengthen duration of syllable *s* by a factor *x* when the next phrase boundary shows proximity *y*"). Some other parameters, though, are applied independent of serial dependency. For example in French, German and English, segment duration is strongly affected by the simple fact that the containing word is either lexical or grammatical in nature.

Overall, current models of this type do not vary a great deal with respect to the variables used for the prediction. In our French speech synthesis, stepwise statistical procedures have identified the following parameters as contributing significantly to segmental duration: speech rate, number of segments in a syllable,

position of the segment in the syllable, lexical or grammatical status of the containing word, position in the word, position in the prosodic phrase, presence or absence of a schwa in the syllable (Keller & Zellner, 1995; Zellner, 1998a). In our current project for the synthesis of Swiss High German (the language used for reading texts aloud in German Switzerland), it was found that this list needed to be extended by just one parameter, lexical stress (Siebenhaar-Rölli, 2000). Similar results are reported in Campbell (1992) for English, and Huber (1991) and Riedi (1998) for German.

Models of this type are found to give relatively satisfactory results, as correlations between 0.85 and 0.9 between predicted and measured syllable durations are typical. Another common measure used to indicate the precision of prediction is the RMSE (root-mean squared error). Values of this measure are currently around 23 ms (Klabbers, 2000). At the same time, perceived timing in speech synthesis systems built with such models still leave something to be desired. This provided the impetus for examining the serial dimension for duration in the current study.

While the $x_1...x_k|x_{k+1}$-type of durational serial dependency examined here is not generally used in timing models, there is some suggestion in the literature that this dimension may be of interest (Gay, 1981; Miller, 1981; Port, 1998; Zellner-Keller & Keller, to appear). While the idea of a *simple syllabic alternation pattern*, such as proposed by by Nishinuma & Duez (1988), is clearly too simplistic a characterisation of a syllable string, it may still hold some merit, in the sense that durational serial dependency may simply be one of the statistically less important contributors to syllable timing, and may thus have been ignored so far. This possibility will be investigated using autocorrelational techniques.

Two theoretically interesting possibilities can be investigated using such techniques. Serial dependency in speech can be seen as occurring either on the linguistic or on the absolute time line. Posed in terms of linguistic (*e.g.*, syllable) time units, the research question is familiar, since it investigates the relationship between syllables that are either adjoining, or situated at lags 2 or more. But the research question can also be formulated in terms of absolute time, by asking about the relationship between elements situated more or less closely from each other in absolute time. Posed in this fashion, the focus of the question is directed more at the cognitive or motoric processing dimension, since it raises issues of neural motor-control or gestural interdependencies within an articulation chain. In this study, both linguistic and absolute time analyses will be performed.

4. Method

Data Bases and Speakers

Two data bases (DBs) served for this study. The first is a 277-sentence data base used for creating a French diphone inventory, the second, a 238-sentence data base used for creating a UK-English diphone inventory. The two DBs had previously been manually segmented using similar criteria by native-language segmentation teams working in or in association with our respective laboratories. Syllable boundaries were assigned on the basis of theoretical criteria, and were corrected, if necessary, with respect to an auditory and visual inspection of the signal. For French, only the first 50 sentences were analysed, for English, the entire 238-sentence DB was analysed.

The French DB contains fast and normal speech rate versions, while the English DB contains only normal speech rate sentences. Speech rates were set by the speaker. "Fast" was defined as "the most rapid speech rate compatible with an errorfree pronunciation", and "normal" was defined as a "speech rate appropriate to a normal, sustained reading of sentence-based text". Average syllable duration for the normal speech rate were 209 ms for French and 206 ms for English, and the average syllable duration of the fast speech rate in French was 149 ms.

The two speakers were male native speakers of the respective language (French and UK English), with no discernible dialectal deviations.

Analysis

The data sets were prepared and analysed using the following steps.

(1) Syllable durations were calculated on the basis of segment durations,

(2) For each half-second of utterance time, the number of syllables was calculated to provide items on the *absolute time line*. For example, if the half second was constituted of two full syllables, plus one quarter syllable at the start, plus one quarter syllable at the end, the value was 2.500. This provided points on the "absolute time line". 500 ms was chosen, since it was observed that no syllable in either DB was ever longer than a half a second.[1]

(3) To obtain comparable numbers for the *linguistic time line*, syllable durations were converted to half-second / syllable duration values, e.g., 500 ms / 206 ms = 2.427. This provided points on the "linguistic time line".

(4) Separate autocorrelations r(k):=corr(x, x_k) were calculated for each utterance. In each utterance and for each lag of 1...k/2, correlation coefficients were calculated for the series created by all pairs x, x_{lag} in the utterance. For example, an utterance extending over ten half-seconds and showing syllable frequencies of 2.775, 2.630, 2.645, 3.559, 2.502, 1.467, 1.671, 3.273, 3.709, 2.255 contains nine pairs with lag 1 (2.775, 2.630...3.709, 2.255), eight pairs with lag 2 (2.755, 2.645...3.273, 2.255), etc. Since the series become very short at k/2, correlations obtained in this fashion vary considerably.

(5) Autocorrelation results were pooled over utterances and identical lags. The means and standard deviations of correlations obtained in this manner are more reliable.

(6) To perform an assessment of statistical significance, we assumed data distribution homogeneity over the pooled data. Significance was assessed with both two-tailed and one-tailed tests. This reflects the wish to test both theoretical possibilities of obtaining either correlations or anti-correlations (two-tailed test), or just anti-correlations (one-tailed test).

[1] Please note that this procedure provides a freely variable time line. For example, it is theoretically possible to calculate a lag of 1.5 by simply counting the number of syllables found in the interval 250-750 ms. An analogous procedure is conceivable for the linguistic time line, where syllable length and frequency can be calculated in more ways than the way it was done here. For this reason, the figures of this study will be present the results as continuous lines.

Fig. 1. French, normal speech rate

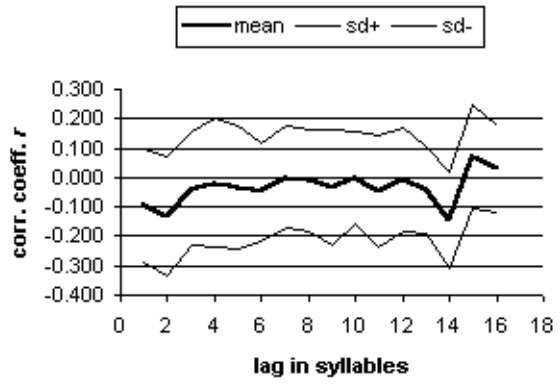


Fig. 2. French, fast speech rate

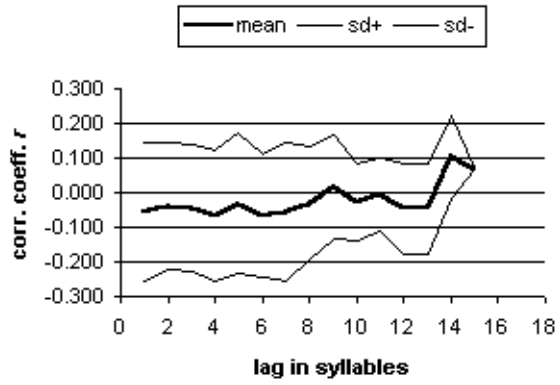


Fig. 3. English, normal speech rate

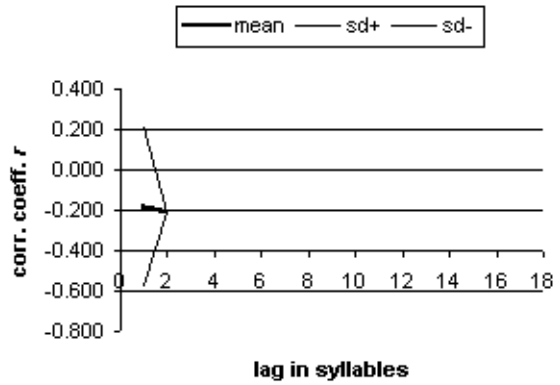


Fig. 4. French, normal speech rate

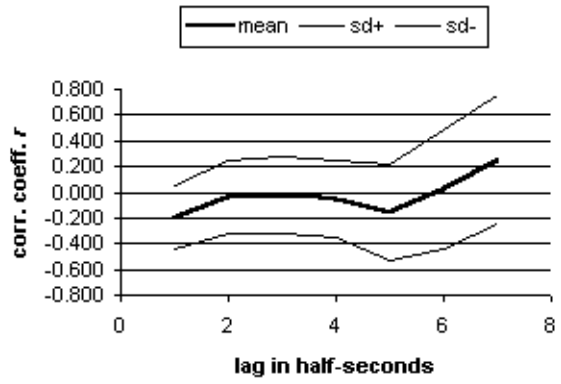
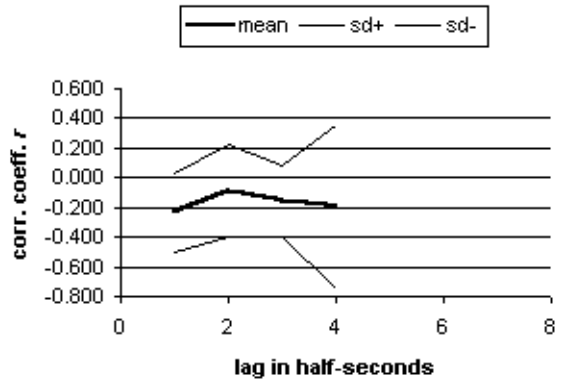


Fig. 5. French, fast speech rate



5. Results

The results are shown in Figures 1 to 5 and Table 1. Figures 1 through 3 show the results for the analysis of the linguistic time line, and Figures 4 and 5 show results for the analysis of the absolute time line.

Generally speaking, it appears that results for the absolute time line were more consistent and "cleaner" than those for the linguistic time line. As judged on the same speech condition, correlation coefficients are stronger, and the evolution of the coefficient over lag is simpler. At a lag of 500 ms, there is a significant anti-

correlation of -0.197 for the normal speech rate in French, and of -0.234 for the fast speech rate in French (English utterances were too short for this analysis). At lags of 1000 ms and more, there were no significant correlations. The second anticorrelation at 2500 ms of lag is suggestive, especially when considered together in conjunction with the results on the linguistic time line, but values did not reach significance, since only 57 observations were made at that lag.

Significant anti-correlations for low lags was also evident in the results for the linguistic time line. Of some theoretical interest is the result of further significant anti-correlations, when tested with one-tailed tests that are specific to anti-correlations, at lags 4 and 6 in the French fast speech condition. If this result can be confirmed in further studies, it might suggest a "ripple effect", whereby negative durational relationships "ripple" down the speech chain and manifest themselves once more 4 and 6 syllables later.

Table 1. Table of autocorrelations

Time Line	Corpus	lag	*r*	*N*	lag	*r*	*N*	lag	*r*	*N*
abs	F, nrm	1	-.197*	389						
abs	F, fast	1	-.234*	226						
ling	F, nrm	1	-.095*	1002	2	-.130*	953			
ling	F, fast	1	-.055**	933	4	-.068**	781	6	-.067**	641
ling	E, nrm	1	-.182*	810						

* significant at p<0.05 two-tailed and p<0.05 one-tailed
** significant only at p<0.05 one-tailed

6. Conclusion

As has been stated frequently, speech rhythm is a very complex phenomenon that relates to an extensive set of predictive parameters. In this study, 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. Also, there is some suggestion of further anti-correlational behaviour at later syllable lags.

Interestingly, anti-correlations measured along the absolute time line are stronger and appear to be more systematic than those measured along the linguistic time line. This relates to the observation that contrary to what a simple syllable alternation model would have suggested, there were no positive correlations to be found at syllable lag 2. Those two observations taken together suggest that the durational serial effects shown here are not simple syllabic effects, but may be related to some cognitive or motoric processing variables.

Indeed, it may be that these speech timing events are subject to a time window roughly 500 ms in duration. Reminiscent of time periods typical of grouped excitatory post-synaptic potentials (EPSPs) as measured intra-neuronally, or corresponding macroscopically potentials measured at the scalp and visible as P300 (Birbaumer et al., 1994), this time period may relate to a neural reduction in cortical excitability, during which antithetical durational effects may be facilitated. Alternative explanations may relate to elastic rebound effects within the articulatory effector system.

Some clues as to the concrete effects of the anti-correlational patterns presented here can be gained by listening to synthetic speech that incorporates a controllable anti-correlational effect over 500 ms of lag time. This has recently been incorporated

into our French-language speech synthesiser (available at www.unil.ch/imm/docs/LAIP/LAIPTTS.html). As judged informally by members of our laboratory, strongly anti-correlational values lend the speech output a pleasant, "swinging" effect, while weak values produce an output that sounds more "regimented". The reading of a news report appeared to be enhanced by the addition of an anti-correlational effect, while a train announcement with strongly anti-correlational values sounded inappropriately "swingy". In that sense, this effect may well relate to a useful control parameter for synthetic speech.

Aknowledgements

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 the University of Lausanne for funding research leaves for the two first authors, hosted in Spring 2000 at the University of York. The study also employs research materials previously funded under a Swiss Fonds National research project and various contracts with BT (UK). Thanks are extended to Prof. F. Bavaud, University of Lausanne, for verifying our statistical procedure.

References

Birbaumer, N., Lutzenberger, W., Elbert, T, Trevorrow, T. (1994). Threshold variations in cortical cell assemblies and behaviour. In H.-J. Heinze, T.-F. M ünte, & G.R. Mangun (Eds.), *Cognitive Electrophysiology* (pp. 248-264).

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.

Gay, T. (1981). Mechanisms in the control of speech rate. *Phonetica 38.* 148 -158.

Huber, K. (1991). *Messung und Modellierung der Segmentdauer für die Synthese deutscher* Z ürich (Dissertation ETH Z ürich 9535).

model for French. *XIIIth International Congress* (pp. 302-305). Stockholm.

fast French. *York Papers in Linguistics* University of York. 53-75. (available at www.unil.ch/imm/docs/LAIP/pdf.files/Keller-

Klabbers, E.A.M. (2000). *Segmental and Prosodic*

Dissertation, Technical University of Eindhoven, NL.

parametric phonetic interpretation and natural-sounding speech synthesis. In Keller, E. (ed).

Recognition. (pp. 253-268). Chichester: Wiley.

Miller, J. L. (1981). Some effects of speaking rate *Phonetica, 38.* 159-180.

perceptive de l'organisation temporelle de l'énoncé en fran çais.
Phonétique d'Aix, 11. 181-201.

Local, J., Carter, P., Dankovicov á J. & Heid, S.. (To appear).
Prosodic Approach to Device-Independent, Natural-Sounding Speech Synthesis. Computer . (see also http://www-users.york.ac.uk/~lang19/york/duration.html)

interpretation in ProSynth, a prosodic speech synthesis system.
International Congress of Phonetic Sciences (Ohala, J.J., Hasegawa, Y., Ohala, M.,

1059-1062. University of California, Berkeley, CA.

Controlling Segmental duration in Speech Synthesis Systems.
ETH. Z ürich. (TIK-Schriftenreihe 26). 174 pp.

segmental durations. In G. Bailly et al., (Ed.). *Talking Machines: Theories, Models, and* (pp. 265 - 273). Elsevier Science Publishers.

The timing model for the Swiss High German LAIPTTS.
Meeting, Stockholm, May 2000. (available at www.unil.ch/imm/docs/LAIP/COST_258/Meeti

pdf).

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

Zellner Keller, B. & Keller, E. (to appear). The

fluency in the language acquisition process. In Ph. Delcloque and V. M. Holland (Eds).

Integrating Speech Technology in Language Learning. Swets & Zeitlinger.

Zellner, B. (1998a). *Caractérisation et prédiction du débit de parole en français. Une étude de cas.* Thèse de Doctorat. Faculté des Lettres, Université de Lausanne. (available at www.unil.ch/imm/docs/LAIP/ps.files/DissertationBZ.ps).

Zellner, B. (1998b). Fast and slow speech rate: A characterisation for French. *ICSLP, 5th International Conference on Spoken Language Processing*. (Volume 7, pp. 3159 - 3163), December 1998, Sydney (Australia).