Comparing Reaction Time Sequences from Human Participants and Computational Models

L. ten Bosch[1], M. Ernestus[12], L. Boves[1]

[1]Radboud University Nijmegen
[2]Max Planck Institute for Psycholinguistics
l.tenbosch, m.ernestus, l.boves@let.ru.nl

Abstract

This paper addresses the question how to compare reaction times computed by a computational model of speech comprehension with observed reaction times by participants. The question is based on the observation that reaction time sequences substantially differ per participant, which raises the issue of how exactly the model is to be assessed. Part of the variation in reaction time sequences is caused by the so-called local speed: the current reaction time correlates to some extent with a number of previous reaction times, due to slowly varying variations in attention, fatigue etc. This paper proposes a method, based on time series analysis, to filter the observed reaction times in order to separate the local speed effects. Results show that after such filtering the between-participant correlations increase as well as the average correlation between participant and model increases. The presented technique provides insights into relevant aspects that are to be taken into account when comparing reaction time sequences.

Index Terms: reaction times, local speed, participant-model comparison, computational modeling, spoken word recognition

1. Introduction

In psycholinguistic experiments, reaction times (RTs) are frequently used as directly observable measures of the mechanisms underlying speech comprehension [1]. RTs are in general straightforward to measure, but difficult to interpret. They are the result of a number of partially concurrent cognitive and execution processes and therefore RT sequences may have a complex structure, as the vast literature shows (e.g. [2]; [3] and references therein, [4]; [5]).

One may expect to be able to better understand the complexity of RT sequences by comparing them with the RT sequences generated by computational models that simulate the underlying cognitive processes e.g. [3, 6, 5]. In this paper we address the question how RT sequences from participants and from a model can be evaluated. In [7], we introduced an end-to-end model of auditory speech comprehension that can simulate a participant in experiments where RT is the main Dependent Variable (DV). We reported low correlations between the RT sequences generated by this model and from human participants. Importantly, however, closer inspection of the human RT sequences showed that the correlations among them are also low.

In this paper, we will investigate the causes of these low correlations. We hypothesize that they result from effects that are independent of the mechanisms underlying the comprehension process. We test this hypothesis by investigating whether filtering of the RT sequences increases the correlations among participants. Moreover, we examine whether filtering also improves the correlations between the participants' RT sequences and those generated by DIANA, an extension of the model described in [7] (cf. Section 4).

RTs are the result of different mechanisms. The effects that best reflect the underlying processing mechanisms result from the specific characteristics of the stimuli. The precise trial-to-trial effect depends on the type of experiment; for example, in lexical decision experiments, the lexical status of the stimulus, its morphological complexity, the density of its lexical neighborhood, and the word or lemma frequency play a role (cf. [8]). DIANA, similar to other models of auditory word comprehension such as Shortlist-B [9] and SpeM [10] only accounts for these stimulus effects.

In addition, participants' RTs are affected by their physical and mental condition, age, gender, handedness, and general cognitive abilities ([11]), which mainly affect the participant's *average* RT in a session [12]. Next, there is a third group of factors that affect RT sequences which are not fixed for the duration of a session but play a role on the intermediate term. These factors become manifest in the form of 'local speed' [13]. The local speed effect is thought to be caused by slowly fluctuating change in e.g. attention, learning effects, and fatigue. Finally, a fourth group of factors, such as a change in a participant's strategy, may cause sudden changes in the local average RT. Not surprisingly, the interplay between all these factors has resulted in several competing approaches in mathematical psychology to modeling RT sequences (e.g. [2, 3, 14, 4, 5, 15] and references therein).

Arguably, the most prominent mechanisms that reduce the correlation between RT sequences from different participants are those that have effect on the intermediate term. This is especially true in experiments in which all participants process stimuli in a different order, since different trial histories lead to different local speed effects influencing the RT on a certain stimulus. We expect the removal of local speed effects to also have an effect on the match between RT sequences from DIANA and from participants, since DIANA does not model local speed and occasional strategy changes.

In this paper we assume that these effects can be captured by an Auto-Regressive-Moving-Average (ARMA) model, the parameters of which can be estimated from the raw RT sequence. Filtering the raw RT sequence by the inverse of that ARMA model should yield an RT sequence that better reflects the short-term, stimulus-dependent reaction times. We prefer ARMA models over models based on the theory of self-organized criticality (SOC) [16, 17, 18] mainly because we do not know how to derive an inverse filter from a SOC model, while deriving an inverse filter from an ARMA model is straightforward. As an aside, we will discuss a relation between

ARMA modeling and mixed effect regression models that are commonly used in the analysis of RTs.

2. Method

We consider the RT sequence obtained in a psycho-linguistic experiment as the superposition of the underlying processing mechanisms, local speed effects, long term effects and participants' strategies. We use an ARMA model to filter the local speed effects from these raw RT sequences.

Estimating the optimal parameters of an ARMA model from noisy data is as much an art as a science. The estimation does not only involve the values of the parameters of the AR and MA parts, but also the estimation of the number of parameters (*order*) of the AR and MA parts. It is known that estimating the parameters of the AR part, given some order, is more robust than estimating the parameters of the MA part. Some approaches to ARMA model estimation divide the process into two steps; the first step estimates the AR parameters, after which the MA parameters are estimated from the error signal that remains after inverse filtering. Other approaches attempt to estimate the AR and MA parameters simultaneously. Parameter estimation is complicated because the additive noise may not be completely white (contrary to the assumption in the mathematical theory) and because the actually underlying process may not be an ARMA model. For example, if a participant's behavior is partly characterized by a SOC model, the ARMA model cannot be completely correct.

In this paper, we focus on estimating the AR part of the filtering process, since this AR part can be directly interpreted as weighted estimation of previous RT values (local speed). This estimation can be performed in two ways: on the entire group of participants, or per participant. More details will be given in section 5. We will show that it is possible to estimate the parameters of an AR model such that, after filtering, the between-participant's correlation as well as the average participant-DIANA correlation increase.

Within the field of psycholinguistics, RTs are often analyzed for the presence of the effect of a certain variable (e.g. a word's frequency of occurrence or semantic transparency). This is mostly done by means of regression analyses, which often incorprate the reaction time on the previous stimulus as a control variable, accounting for the effect of local speed (cf. [13, 19]). We will briefly discuss how ARMA modeling may also improve this analysis of RTs. The ARMA filter represents more noise resulting from local speed than is captured by the single use of the RT on the previous trial. In fact, the previous RT can be seen as a special case of an ARMA filter.

3. Word recognition experiment

For assessing the RTs from participants and DIANA, we conducted a word recognition experiment.

3.1. Participants and materials

Twenty native listeners (10 male, 10 female, 18 to 23 years) without reported hearing problems were paid to participate in this recognition experiment.

The stimuli of this experiment consist of 613 Dutch real words. All words were chosen from the Spoken Dutch Corpus (CGN) dictionary ([20]). They have single bi-syllabic stems, and include plurals of nouns, inflected forms of adjectives, past tense and past participle forms of verbs. The list contains words of very low, medium and high frequency, with a bell-shaped distribution on the log-frequency axis.

To obtain the auditory stimuli, all 613 words were carefully read aloud by one single female native speaker. Words were spoken in isolation. The duration of the words varies from 273 to 947 ms, with a mean of 552 ms. For each participant, a list was created such that all stimuli could be presented in a random, participant-specific ordering.

3.2. Procedure

Participants were asked to press a button as soon as they thought they had recognized the word. They then had to repeat the word; this could be done without time pressure, after the button was pressed. The responses were recorded and used to compute the proportion correct responses; during the experiment no feedback about correctness was presented to the participant.

Stimuli were presented via headphones, and the experiment took place in a sound-attenuated room. The button box used was connected to a dedicated stand-alone PC with E-prime as single main process. The auditory presentation immediately stopped at the moment the button was pressed. The list of 613 stimuli was split into four sublists and participants were offered the opportunity to take a short rest between sublists. One experimental session (covering the entire list of 613 stimuli) took approximately 50 minutes.

4. DIANA: A computational model of speech comprehension

DIANA, the computational model to be assessed in this paper, simulates participants' behavior in experiments in the field of spoken word comprehension. DIANA has much in common with the model described in [7]. In line with the architecture common to many other models (see e.g. [11]), DIANA consists of three components: a word activation component, a decision component, and an execution (effector) component. Contrary to the model described in [7], where activation and decision formed a pipeline, in DIANA activation and decision operate in parallel. DIANA is an end-to-end model, which means that it simulates the entire processing from the acoustic input up to the key press and the output of the word that was recognized. The time between the onset of the word and the button press is the RT, measured in real time.

The activation component takes as its input the acoustic signal, unfolding over time, and computes the activation of internal word representations that are stored in DIANA's lexicon. Its implementation is based on HTK [21]. The activations, which are updated each 10ms, are given by

$$\log P(\text{signal}|\text{word}) + \lambda \log(P(\text{word})) \qquad (1)$$

λ governs the balance between the bottom-up acoustic information (first term) and the top-down linguistic information (second term). For this paper, we determined the value for λ ($\lambda = 2.5$) by optimizing DIANA's recognition accuracy on a different set of words produced by the same speaker.

DIANA's activation component builds a time-varying ranked list of word hypotheses which is constructed in DIANA's word search space. For each time t between stimulus onset and stimulus offset, this ranked hypothesis list is accessible for DIANA's decision component, which is tightly coupled with the activation component. Here, a decision about the winning word hypothesis is made on the basis of whether the activation of a provisional winning word hypothesis at time t exceeds the

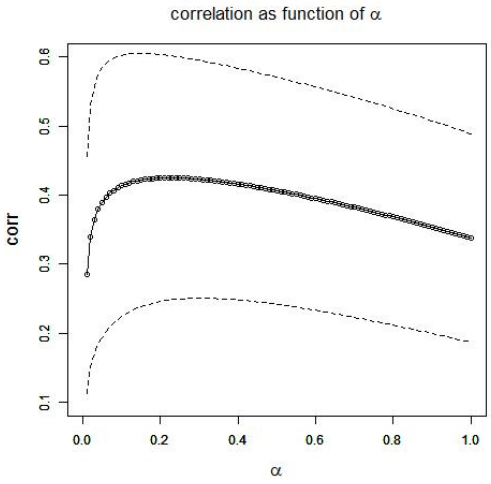


Figure 1: *Correlation between the raw and filtered RT sequences as a function of* α*. The solid line shows the correlation by applying one filter optimized on the entire group of participants; the dashed lines indicate the standard deviation from the mean of all correlations on the basis of individual participants. The value* $\alpha = 1$ *always is equal to an effect of only the previous RT.*

activations of all competing words with a specified threshold θ. This decision (which takes place at time RT_{dec}) initiates DIANA's execution component. In the current simulation, DIANA's execution component is assumed to simply add RT_{exe} (in the current simulations fixed to 100ms) to simulate the execution time between the mental decision and the overt behaviour (key press). Eventually, DIANA's output consists of a triplet: a reaction time [RT_{dec}+ RT_{exe}], the word that was recognized (the winner at the moment of decision), and the total activation score (determined by the activation component on the basis of the match between signal and internal representations).

The computed RT depends on the stimulus, on the settings in DIANA's activation component (the underlying acoustic model, the value of λ, the prior probabilities of the words in the language model), and on θ. For the experiment reported here, the settings of the activation component are all fixed. The threshold θ governs the long-term effects in the simulated RT sequence: the smaller θ, the smaller RT_{dec} (and the greater the risk that a response is incorrect). For this paper we used a number of different values for θ, but its value was always fixed during a simulation run. Its eventual value was chosen on the basis of the optimization of correlations between the humans' RTs and those generated by the model (see section 5).

5. Analysis and Results

Prior to filtering, implausible RTs in the participant data (faster than 200ms after word onset and slower than $\mu + 2\sigma$ after word onset, where μ and σ were determined on the complete sequence produced by one participant) were excluded. The resulting raw RT sequences are depicted in Fig. 2, left panel. The figure clearly shows an enormous difference between participants, both across and within the course of the experiment.

Next, the local speed filtering was done. We applied an AR filter, parametrized by one parameter α, on the raw RTs, and investigated for which α the correlation between the raw and the filtered RT sequences was maximal in order to optimally

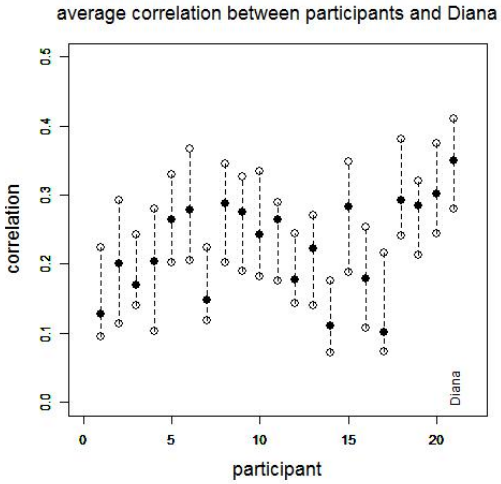
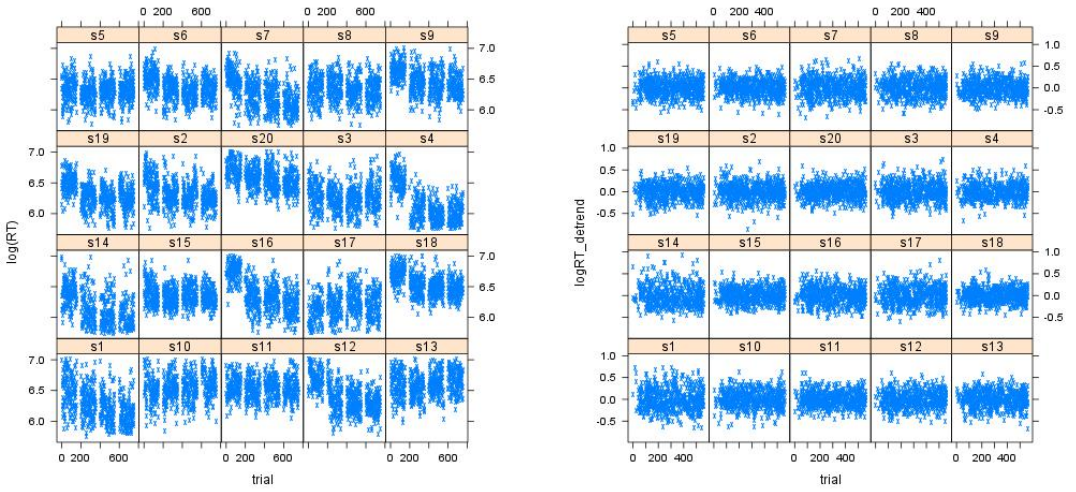


Figure 3: *The horizontal axis presents participants, with* DIANA *as number 21. The vertical axis displays correlation. The figure displays three cases: no filtering (the lower open circles), group-based filtering (closed circles) and individual filtering (upper open circles).*

capture local speed effects. Fig. 1 shows the average correlation between the raw and AR-processed RTs across participants as a function of α, with an optimum around 0.19-0.21. This value corresponds to an average history of the $1/\alpha \approx 5$ recent RT values required to optimally predict the current RT. This is in agreement with findings reported in other reaction time studies where typical ranges are 5–10 ([15], p. 409). The left and right panel in Fig. 2 show (per participant) the raw RT data and the RT data after this local speed filtering, respectively.

Since removal of local speed effects amounts to reducing the effects of factors that are not directly related to spoken word recognition proper, it is to be expected that the correlations between the individual participants will increase after subtracting the estimated local speed (as obtained from the AR analysis) from the raw RTs. Moreover, because DIANA does not model the effect of local speed at all, the correlations between DIANA and human participants are likely to increase as well. That this is indeed the case is shown in Fig. 3. The figure presents the average correlation between each participant with other participants (columns 1 to 20) and the average correlation between DIANA (as a virtual participant) and all human participants (the right-most column). The lower open circles represent the correlations obtained without filtering. In this case, the average correlation between the RT sequences of any pair of participants is low ($r = 0.16$); the average human-DIANA correlation is slightly higher ($r = 0.28$). The closed circles correspond to the situation after filtering; the correlations increase to $r = 0.22$ and $r = 0.35$, respectively. The eventual value of θ in DIANA's decision component was chosen to optimize these human-DIANA correlations. Clearly, filtering increases the average correlation between human RTs as well as the average human-DIANA correlations.

In all cases, the human-DIANA correlation exceeds the average human-human correlation. Since DIANA only simulates processes that are directly involved in spoken word recognition and decision making, this suggests that DIANA is able to capture, at least to a large extent, the effects from spoken word recognition processes as used by the participants in generating their RT sequences.



(a) Raw RT sequences

(b) Filtered RT sequences

Figure 2: *RTs (vertical axes) before filtering (left) and after inverse filtering with an AR-1 model (right), as function of stimulus index (horizontal axes). Each cell represents a participant.*

These correlations can be further improved (to $r = 0.29$ between participants, and to $r = 0.41$ for participants versus DIANA) when the parameter of the AR filter is estimated *per participant* instead of on the entire group (shown by the upper open circles in Fig. 3). It appears that the optimal values of α differ substantially across participants; they range from about 0.11 to 0.80, showing substantial differences between participants with respect to their estimated local speed.

The value $\alpha \approx 0.2$ is interesting because it reappears when the AR filter is incorporated in regression models predicting raw RTs. We observed this by contrasting pairs of regression models that all share the predictors 'duration of stimulus', 'log word frequency' as fixed predictors, and stimulus and participant as random effects, but pairwise only differ in the way the AR filter is incorporated: as fixed effect (i.e. participant independent) or as random slope under participant (participant dependent). In this way, six different model pairs were compared (by varying the presence versus absence of interactions). Consistently the AIC of the model with the AR filter as random slope is lower than the AIC of the alternative model with the AR filter as fixed effect only, but only if $\alpha > 0.17$. This shows again that while almost all participants' RTs are correlated with the preceding two, three RTs, some participants show much longer term local speed effects. In addition, the analysis shows that the AIC of these regression models can be improved by replacing the predictor 'previous RT' by a proper AR filter.

6. Discussion and Conclusion

RT sequences are relatively easy to obtain, but are intrinsically complex to model. They are influenced by factors at several different levels, each with a different time domain. We explored ways for estimating an AR filter (e.g. [12]) that removes part of the effects on RT sequences other than those directly related to stimuli (and thus the underlying speech comprehension process). The results showed that this filtering significantly increases both the correlations between individual participants and between participants and DIANA, our end-to-end model of speech comprehension. This strongly suggests that DIANA is able to simulate RTs as if they were coming from a cognitive process that underlies the RTs from all participants. The time domain of the local speed effect as estimated for the entire group is 5 stimuli, but this domain largely varies across participants. This might of course call into question how much sense it makes to estimate the time domain on the basis of all participants' RTs. Indeed, although Fig. 3 shows that correlations between participants and DIANA increase on the basis of group-based estimations of this domain, the best improvements can be achieved by proper participant-dependent analysis.

The analysis in this paper may have impact on how local speed is dealt with in regression models analysing the effect of specific predictors on RTs. Our analyses show that local speed is better captured by an AR filter with $\alpha = 0.2$ than with RT to the previous stimulus as predictor (which is in fact an AR filter with $\alpha = 1$). Regression models using an AR or ARMA filter as predictor result in better fits with the data (and lower AICs).

In DIANA we replaced the Linear Ballistic Accumulator model diffusion model of the decision process (e.g. [5, 22]), which we used in the predecessor model in [7], by a simple threshold on the distance between the activations of the top word hypothesis and the runner-up. While we believe that activation and decision making are indeed parallel, rather than sequential processes, future research is necessary to see whether some features of diffusion models can be integrated in DIANA and whether this will make it possible to increase even further the correlation between the RTs generated by the model and by human participants in a principled way.

In conclusion, this paper has shown that human participants show low correlations among their RTs in speech comprehension tasks due to local speed effects (e.g. learning, fatigue, etc.). Moreover, due to these effects, computational models of speech comprehension, which do not incorporate these effects, cannot show high correlations with participants' behavior. Therefore, the effects of the underlying processing mechanisms can only be well investigated after the local speed effect has been partialled out per participant (for instance, with an ARMA filter).

7. Acknowledgement

This work was funded by an ERC starting grant (284108) and an NWO VICI grant awarded to Mirjam Ernestus.

8. References

[1] R. Whelan, "Effective analysis of reaction time data," *The Psychological Record*, vol. 58, no. 3, pp. 475 – 483, 2008.

[2] A. Kelly, A. Heathcote, R. Heath, and M. Longstaff, "Response time dynamics: evidence for linear and low-dimensional nonlinear structure in human choice sequences," *The quarterly journal of Experimental Psychology Section A: Human Experimental Psychology*, vol. 54, no. 3, pp. 805–840, 2001.

[3] D. E. Meyer and D. E. Kieras, "A computational theory of executive cognitive processes and multiple-task performance:part 1. basic mechanisms," *Psychological Review*, vol. 104, no. 1, pp. 3–65, 1997.

[4] R.Ratcliff and J. N. Rouder, "Modelling response times for two-choice decisions," *Psychological Science*, vol. 9, p. 347, 1998.

[5] S. Brown and A. Heathcote, "The simplest complete model of choice response time: Linear Ballistic Accumulation," *Cognitive Psychology*, pp. 153–178, 2008.

[6] E. Wagenmakers, S. Farrell, and R. Ratcliff, "Estimation and interpretation of $1/f^{\alpha}$ noise in human cognition," *Psychonomic Bulletin & Review*, vol. 11, pp. 579 – 615, 2004.

[7] L. ten Bosch, L. Boves, and M. Ernestus, "Towards an end-to-end computational model of speech comprehension: Simulating a lexical decision task," in *Proceedings of Interspeech*, Lyon, France, 2013.

[8] A. Cutler, *Native Listening: Language Experience and the Recognition of Spoken Words*. MIT Press, 2012.

[9] D. Norris and J. McQueen, "Shortlist B: A Bayesian model of continuous speech recognition," *Psychological Review*, vol. 115, pp. 357 – 395, 2008.

[10] O. Scharenborg, D. Norris, L. ten Bosch, and J. McQueen, "How should a speech recognizer work?" *Cognitive Science*, vol. 29, pp. 867 – 918, 2005.

[11] J. J. Lee and C. F. Chabris, "General cognitive ability and the psychological refractory period: Individual differences in the minds bottleneck," *Psychological Science*, vol. 24, no. 7, pp. 1226 – 1233, 2013.

[12] E. Wagenmakers, S. Farrell, and R. Ratcliff, "Estimation and interpretation of $1/f^{\alpha}$ noise in human cognition," *Psychonomic Bulletin & Review*, vol. 11, pp. 579 – 615, 2004.

[13] M. Ernestus and R. H. Baayen, "The comprehension of acoustically reduced morphologicaly complex words: the roles of deletion, duration, and frequency of occurrence," in *Proceedings of ICPhS*, Saarbrücken, 2013, pp. 773–776.

[14] R. Ratcliff, "Group reaction time distributions and an analysis of distribution statistics," *Psychological Bulletin*, vol. 86, pp. 446 – 461, 1979.

[15] T. L. Thornton and D. L. Gilden, "Provenance of correlations in psychological data," *Psychonomic Bulletin & Review*, vol. 12, pp. 409 – 441, 2005.

[16] P. Bak, *How nature works: The science of self-organized criticality*. New York: Springer-Verlag, 1996.

[17] G. C. Van Orden, J. G. Holden, and M. T. Turvey, "Selforganization of cognitive performance," *Journal of Experimental Psychology: General*, vol. 132, pp. 331 – 350, 2003.

[18] G. Van Orden, J. Holden, and M. Turvey, "Human cognition and 1/f scaling," *Journal of Experimental Psychology: General*, vol. 134, pp. 117 – 123, 2005.

[19] I. Hanique, E. Aalders, and E. Ernestus, "The robustness of exemplar effects in word comprehension," *The Mental Lexicon*, in press.

[20] N. Oostdijk, "The spoken dutch corpus project," *ELRA newsletter*, vol. 5, no. 2, pp. 4–8, 2000.

[21] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK book (for HTK version 3.4)," Cambridge University Engineering Department, Cambridge, UK, Tech. Rep., 2009.

[22] C. Donkin, L. Averell, S. Brown, and A. Heathcote, "Getting more from accuracy and response time data: Methods for fitting the Linear Ballistic Accumulator model," *Behavior Research Methods*, vol. 41, pp. 1095–1110, 2009.