Interspeech 2018
2-6 September 2018, Hyderabad

Analyzing Reaction Time Sequences from Human Participants in Auditory Experiments

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

[1]Radboud University Nijmegen, NL; [2]Max Planck Institute for Psycholinguistics

l.tenbosch, m.ernestus, l.boves@let.ru.nl

Abstract

Sequences of reaction times (RT) produced by participants in an experiment are not only influenced by the stimuli, but by many other factors as well, including fatigue, attention, experience, IQ, handedness, etc. These confounding factors result in long-term effects (such as a participant's overall reaction capability) and in short- and medium-time fluctuations in RTs (often referred to as 'local speed effects'). Because stimuli are usually presented in a random sequence different for each participant, local speed effects affect the underlying 'true' RTs of specific trials in different ways across participants. To be able to focus statistical analysis on the effects of the cognitive process under study, it is necessary to reduce the effect of confounding factors as much as possible. In this paper we propose and compare techniques and criteria for doing so, with focus on reducing ('filtering') the local speed effects. We show that filtering matters substantially for the significance analyses of predictors in linear mixed effect regression models. The performance of filtering is assessed by the average between-participant correlation between filtered RT sequences, and by Akaike's Information Criterion, an important measure of the goodness-of-fit of linear mixed effect regression models.

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

1. Introduction

In psycholinguistic experiments, reaction times (RTs) are frequently used as observable measures of the cognitive effort necessary to complete a task, e.g. [1]. RTs are easy to measure. At the same time, however, RTs are difficult to interpret, if only because observed RTs (e.g., measured via a button press) are the combined result of several sequential and parallel cognitive, neuro-physiological and mechanical processes (see, e.g., [2, 3]), each with different effects on the observed RT.

One short-term effect is related to the stimulus itself (its lexical status, phonetic make-up, morphological complexity, the density of its lexical neighborhood, frequency, indexical effects, etc.), but RTs are also affected by factors that are not related to the stimuli. These confounding factors include long-term effects (participant's health condition, age, gender, handedness, general cognitive abilities, gaming experience, etc. [4]) and medium-term effects (attention fluctuation, strategy changes, fatigue) (see, e.g., [5, 6, 7, 8, 9, 10, 11, 12] and references therein). These medium-term effects are collectively referred to as 'local speed effects'.

In experimental set-ups, participants are usually exposed to randomized trial ordering. Local speed effects make it difficult (or senseless) to compare raw RTs responded on the same trial across participants. Indeed, the between-participant correlations between observed RT sequences are typically low, sometimes insignificant, and sometimes even significant but negative (e.g. [13, 14]). One way to analyze and understand RT sequences is by comparing them to computational model simulations, e.g. [7, 10, 11, 15, 13, 14, 3]. In this paper we focus on the RT sequences as observed. We present a number of options for filtering the long-term and medium-term effects from the observed RT sequences, and show the large impact of filtering on the analysis and interpretation of linear fixed effects regression (lmer, [16]). Linear mixed effect models are often used to analyze RT data and to investigate significance levels of predictors of interest ([17, 18], see also [19]) in auditory lexical decision and word comprehension experiments. Typically, lmer models have the RT given to the previous stimulus ('previous RT'), which captures the local speed effects, as significant predictor. In this paper, we show the consequences of replacing 'previous RT' by a filtered (detrended) variant.

Filtering can be performed in many ways. Our first assessment criterion does not require any regression model, but instead is purely based on the average between-participant correlation between filtered RT sequences. As a second assessment criterion, we use the Akaike Information Criterion (AIC) of lmer models in which the filtered RT replaces the 'previous RT' as one of the predictors.

2. Data

For this research we re-used the RT data in the BALDEY corpus [20]. Twenty native listeners (10 male, 10 female, 18 to 23 years) without reported hearing problems were paid to participate in this lexical decision experiment. For each of the 20 participants, the experiment consisted of 10 sessions, one per week. Each participant made lexicality decisions on a total of 5541 stimuli, about half of which were pseudo words. In the analyses in this paper we used the log-transformed RT sequences recorded in the individual BALDEY sessions.

3. Method

Our detrending method is based on the idea to consider a sequence of RT values as a signal varying over time, with low, mid and high frequency components. If we assume that an observed RT sequence can be modeled as a superposition of high-frequency effects imposed by individual stimuli and medium- and low-frequency effects that can be subsumed under 'local speed', removing these local speed effects boils down to applying a high-pass filter with a adequately chosen cut-off frequency. For example, the RT on the preceding stimulus can be considered as a high frequency component, and as a point estimate of the amplitude of the local speed wave. As with any point estimate, this may be quite unreliable, but it does have the advantage that its effect is very local and confined to a single stimulus – if it turns out to be an outlier, its impact is limited.

10.21437/Interspeech.2018-1728

Several techniques are available for obtaining filters that are scalable and that generalize beyond point estimations. In this paper we propose three options (see section 3.2 and further).

3.1. Filtering and linear mixed effect models

A frequently used lmer model that predicts the RT sequences in psycholinguistic experiments uses fixed effect predictors such as word frequency, the duration of the stimulus, the previous RT, information about the morphology of the stimulus, and subject and stimulus as random effects, with previous RT as random slope, e.g., [18]:

```
lmer_model = lmer(log(RT) ~ log(wordfreq)
+ log(dur) + ... + log(RTprev)
+ (1|subject) + (1 + log(RTprev) | stim))
```

Invariably, word duration (dur) and the 'previous RT' (RTprev) are among the most significant predictors of the current RT. [13, 14] proposed to replace the 'previous RT' with a weighted average of a larger number of preceding RT values. This weighted average is denoted maRT = ma$RT[1, 2, \ldots, N]$, where N denotes the number of RT measurements in the experiment session to be detrended. In their proposal the effect of previous RT values decays exponentially with recency:

$$\begin{aligned} \text{ma}RT[1] &= RT[1] \\ \text{ma}RT[i] &= \alpha RT[i-1] + (1-\alpha)\text{ma}RT[i-1] \quad (1) \\ &\quad \text{for } i > 1, \end{aligned}$$

with $0 < \alpha \leq 1$. The resulting ma$RT[i]$ ('moving average RT') serves as generalization of 'previous RT' in lmer models. Observe that maRT generalizes the 'previous RT': by setting $\alpha = 1$ we see that 'previous RT' is a special case.

In terms of filters, maRT is the output of a low pass filter in which α determines the cut-off frequency. In the detrended RT sequence $RT - \text{ma}RT$, long-term effects and mid-term effects that were present in RT have been removed. This can easily be seen in the case when the RT sequence consists of the same RT values: the maRT sequence will converge to the RT sequence and the difference sequence will therefore approximate 0.

It can be shown that the 'length' of the relevant RT history in maRT can be approximated by $\approx 1/\alpha$ stimuli. Since typical lengths for the relevant RT history vary between 5 and 10 ([12], p. 409), values of α around 0.1 can be expected to be optimal.

3.2. Participant-independent α

We applied detrending using Eq. 1 on the RT data in BALDEY [20], and investigated the impact of the detrending parameter α on the average correlation between the detrended log-RT sequences across all pairs of participants. Ideally, a proper detrending removes the participant's individual local speed effects and therefore increases the average correlation between participants. Given the observed RT sequence, outliers are identified and log() is applied, the sequence is detrended (as a function of α), and finally resorted to keep in line with the same stimulus ordering for all participants. The average correlation between the 190/2 pairs of two participants as a function of α is shown in Figure 1. The horizontal axis displays α; the vertical axis shows the average between-participant correlation. It can be seen that the average correlation has a fairly narrow maximum, and that the optimum correlation (0.174) is reached for $\alpha \approx 0.05$. This implies that in BALDEY the RTs of about

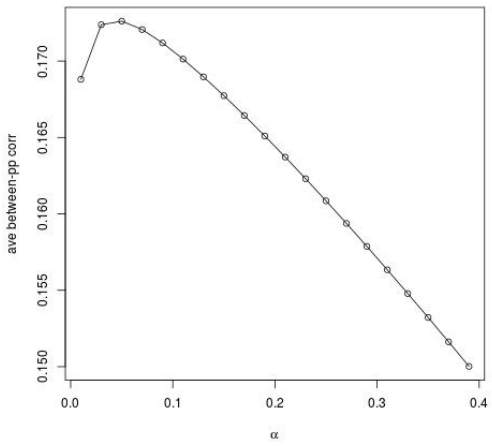

Figure 1: *Average correlation between resorted detrended log(RT)-sequences in BALDEY, averaged over all between-participant pairs, as a function of the detrending parameter α.*

20 previous trials are needed to obtain a stable estimate of the current local speed effect.

This result can be considered in parallel with an AIC analysis. Fig. 2 displays the AIC of three different regression models as a function of α. These three models are variants of the regression model shown above, in which 'previous RT' is replaced by maRT. In an exploratory analysis, i.e. the search for a good regression model that 'explains' the observed data, one often has to contrast subtly different lmer models, and the significance of predictors of interest is directly related to the competition between lmer models (e.g., [19]). The models chosen in this experiment differ with respect to the inclusion of the predictor log(word frequency) (absent in 1, present in 2 and 3) and the presence of random slope maRT under 'stimulus' (absent in 1 and 2, present in 3).

The AIC of all these models appear highly dependent on α. Fig. 2 clearly shows the same trend as the between-pp correlation. Overall, the impact on AIC of changing α is much larger than the impact of changing the predictor structure in the regression models.

In Fig. 2, the relative *ordering* of the models seems independent of α. If that were true, the α-dependency of the AIC would be harmless for significance analyses. This is, however, not necessarily true. Clearly the inclusion of log frequency (model 1 versus 2, 3) provides a significant improvement, independent of α. However, when we focus on model 2 and 3 in Fig. 2 we obtain Fig. 3, showing that AIC(model 2) $-$ AIC(model 3) changes sign about halfway the α interval. The vertical gray bar indicates the α region where the absolute value of the AIC difference is smaller than 2, that is, where the models do not significantly differ (Akaike criterion). As a consequence, the value of α determines the statistical model to be preferred; in this case, whether or not maRT is a random slope under stimulus or not, in other words, whether local trends matter in the RTs per stimulus, or not.

3.3. Participant-dependent α

The improvements in the previous section were based on a detrending with a group-wide value of α. In the second experiment, we allow the value of α to be participant-dependent. It is

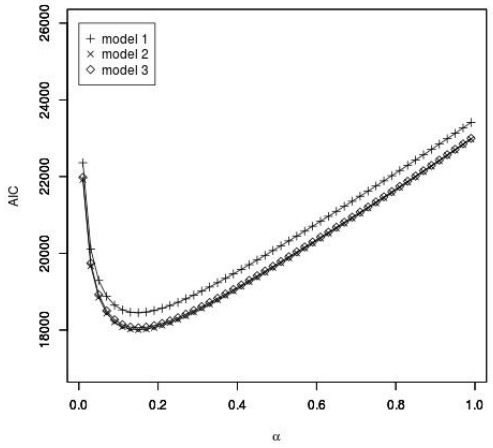

Figure 2: *AIC values of 3 different regression models on BALDEY data as a function of detrending parameter α.*

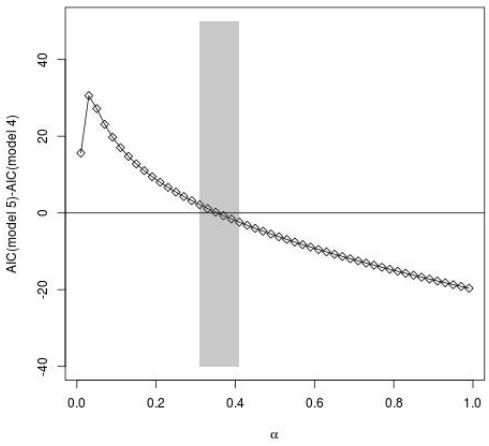

Figure 3: *Difference between the AIC values of two competing regression models 2 and 3 in Fig. 2 as function of α. The gray bar indicates the α region there the absolute value of the difference is smaller than 2.*

assumed constant across the participant's sessions in BALDEY. For the optimization of the average between-participant correlation, the 20 optimal values of α are determined via a multivariate optimization function (nlm(), in R [21]).

The resulting 20 values of α were all between 0.05 and 0.1, except for two participants (with $\alpha \approx 0.4$, and $\alpha \approx 0.01$). A high α means that the current RT is influenced substantially by only the previous RT (that is, the local speed effects are really local); a value of α close to 0 means that the current RT is better predicted by a longer RT history.

Interestingly, the participant-dependent detrending only leads to a non-significant and minor improvement (r=0.177) of the average between-participant correlation, compared to group-wide α (r=0.174). Closer analysis shows that participant-dependent detrending helps substantially to improve the correlation between two participants (to r=0.2, approximately), to a smaller degree for a group of 5 participants, and hardly for a group of 20 participants. The same holds for the AIC of the regression models: improvements for participant-dependent are clear for a small number of participants, but are completely washed out in case of 20 participants.

3.4. Dynamic detrending

It is questionable whether a fixed low pass filter can cope with dynamic changes in local speed. Therefore, we conducted an experiment in which we allow α to dynamically change over time. This is done by

$$\alpha = \arg\min \sum_i (RT[i] - \mathrm{ma}RT[i])^2 \qquad (2)$$

in which i denotes the trial index, and the sum is taken over a window of 100 consecutive RT values, instead of over the entire history. By sliding this window along the entire RT sequence, one obtains an estimation of the locally best value of α. (In the equation, maRT depends on α via eq. 1.)

Figure 4 shows the resulting dynamic detrending parameter α as a function of trial index for the first participant in a session halfway the BALDEY experiment. The figure shows the participant's adaptation directly from the start of the session, after which a stable pattern is reached for a short time; the stable trend is often disrupted again later in the session.

Figure 5 shows an overlay of all sessions of participant 1. This participant has a clear session-related behavior. On top of this trend, there are local disruptions, especially in the second half of the sessions, probably related to fatigue or distraction.
Table 1 shows the average between-participant correlation without detrending, by using the conventional 'previous RT'-based detrending, the participant independent detrending, the participant dependent detrending, and the dynamic detrending.

Table 1: *Average between-participant correlation values for different detrending options.*

no detrending	0.06 ± 0.01
detrending by 'RT previous'	0.11 ± 0.01
participant independent detrending	0.17 ± 0.01
participant dependent detrending	0.17 ± 0.01
dynamic detrending	0.19 ± 0.01

4. Discussion and Conclusion

The results in this paper clearly show that RTs are a complex phenomenon. By using a detrending procedure on BALDEY data, we are able to generate a detrended version of the raw RT sequence that extends the use of the 'previous RT' in lmer models. By a proper choice of α, it is possible to both improve the between-participant correlation between RT sequences as well as the AIC of many lmer regression models. The resulting values of α (here 0.05 and 0.18) are not the same, but the fact that they are much smaller than 1 shows that the complex temporal structure in BALDEY is only partially captured by lmer models using 'previous RT'.

Detrending matters for significance analyses. In minimally differing regression models that are in close competition with respect to their AIC values, a change in α may change the statistical model to be preferred and may turn a predictor deemed

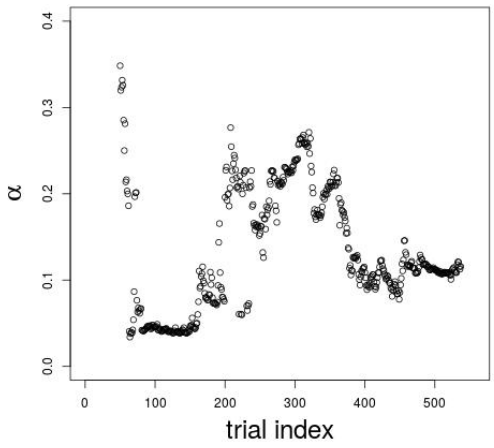

Figure 4: *Dynamic α as function of trial index, for participant 1 in BALDEY session 10 (i.e., halfway the total experiment).*

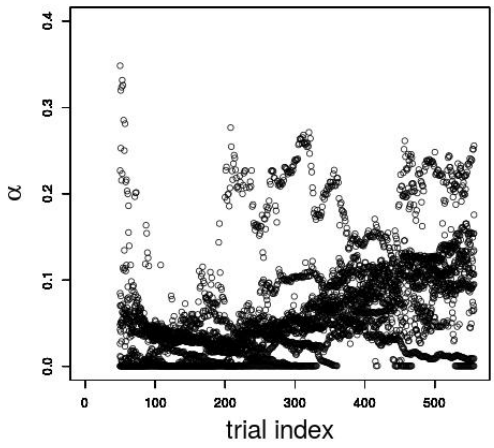

Figure 5: *An overlay of the dynamic α across all sessions of one participant, showing that although there is a global behavioral trend visible across sessions, many local mid-term interruptions and changes of behavior may occur during a session.*

significant to become insignificant and vice versa. This reasoning holds in general, taking into account the recent debates about the exact computation of the number of degrees of freedom in an lmer model, and the applicability of AIC in lmer models, especially with complex random structure [22, 23, 24].

Detrending has a clear interpretation in terms of the spectral properties of RT sequences. Invariably, researchers make sure that the random stimulus sequences they create are random, in all features that are under experimental control. As a result, it may be expected that all stimulus duration sequences are random, meaning that, taken as a time series, they are expected to have a flat spectrum. Figure 6 shows the average spectrum of 1000 different random orderings of the stimulus durations in the experiment reported in [15]. The figure confirms the fact that the expected spectrum is flat.

The mirror-image correspondence between Figs 1 and 2 may be rather surprising, since the measured used are very different. Mathematically, both methods deal with minimization of squared differences between observed and predicted RT data, but they differ in the way how the variance is explained: in terms of stimuli and participant properties, or in terms of correlations between weighted RTs over a long RT history.

The second experiment, in which α was participant-dependent, did not bring any improvement in terms of a better between-participant correlation, nor in terms of AIC of regression models. Apparently, a participant-focused detrending is not necessarily improving the performance on a group of participants when the group size exceeds 10 pps. The lack of improvement in this case might also be due to the fact that the 10 BALDEY sessions per participant were often quite different in their RT patterning.

The third experiment shows that it is possible to detect changes in participant's behavior over experiment sessions. In the beginning of each sessions, the participant clearly has to adapt. The second half of most sessions is characterized by an increase of disruptions in the RT patterning, probably related to fatigue and changes in attention. In addition, there are substantial differences between sessions.

In the near future, we plan to experiment with alternative and fundamentally more profound methods for trend removal, by e.g. applying Chebyshev polynomials to the sequences of log-RTs. Chebyshev polynomials are preferred over regular

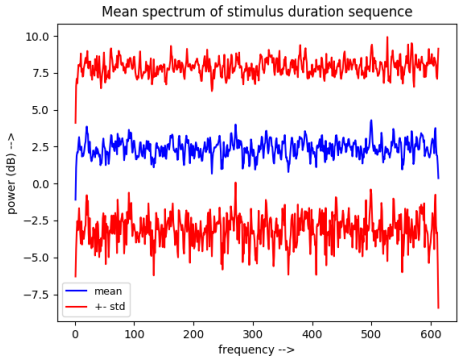

Figure 6: *Average spectrum of multiple reorderings of the stimulus durations, on BALDEY.*

polynomials we were using in this paper because the former are less ill-behaved outside the interval of observation. We expect that non-linear trend removal with a Chebyshev polynomial of a reasonable order does not remove all seemingly structured medium-term fluctuations from the log-RT sequences. An inverse filtering operation, which is known from speech processing, can then be applied for additional spectral flattening. This could be achieved by applying a covariance LPC analysis [25] to overlapping sequences of log-RT observations, and by using the predictor parameters as the coefficients in an inverse filter. The combination of such a detrending method with linear mixed effect modeling is topic for further research. Another, related future direction is the deepening of the connection with Generalized Additive Models (e.g., [26]), spectral approaches (e.g., [10]) and risk modeling (e.g., [27]).

5. Acknowledgement

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

6. References

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

[2] H. Baayen and P. Milin, "Analyzing Reaction Times," *International Journal of Psychological Research*, vol. 3, no. 2, 2010.

[3] S. Sternberg and B. Backus, "Sequential processes and the shapes of reaction time distributions," *Psychological Review*, vol. 122, no. 4, pp. 830–837, 2015.

[4] 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.

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

[6] 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.

[7] 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.

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

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

[10] 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.

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

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

[13] L. ten Bosch, L. Boves, and M. Ernestus, "Comparing reaction time sequences from human participants and computational models," in *Proceedings of Interspeech*, Singapore, 2014.

[14] ——, "DIANA: towards computational modeling reaction times in lexical decision in North American English," in *Proceedings of Interspeech*, Dresden, 2015.

[15] ——, "Towards an end-to-end computational model of speech comprehension: Simulating a lexical decision task," in *Proceedings of Interspeech*, Lyon, France, 2013.

[16] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013. [Online]. Available: http://www.R-project.org

[17] 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.

[18] I. Hanique, E. Aalders, and M. Ernestus, "How robust are exemplar effects in word comprehension?" *The Mental Lexicon*, vol. 8, no. 3, pp. 269–294, 2013.

[19] R. H. Baayen, J. van Rij, C. de Cat, and S. N. Wood, "Autocorrelated errors in experimental data in the language sciences: Some solutions offered by Generalized Additive Mixed Models," in *Mixed Effects Regression Models in Linguistics*, D. Speelman, K. Heylen, and D. Geeraerts, Eds. Berlin: Springer, (to appear).

[20] M. Ernestus and A. Cutler, "BALDEY: A database of auditory lexical decisions," *Quarterly Journal of Experimental Psychology*, vol. Advance online publication, 2015.

[21] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org

[22] M. Baker, "Statisticians issue warning on P value," *Nature*, vol. 531, p. 151, 2016.

[23] D. Li, 2015. [Online]. Available: https://daijiang.name/en/2015/06/22/why-no-p-values-in-mixed-models/

[24] H. Liang, H. Wu, and G. Zou, "A note on conditional AIC for linear mixed-effects models," *Biometrika*, vol. 95, no. 3, pp. 773–778, 2008.

[25] J. Markel and A. Gray, Jr., *Linear Prediction of Speech*. Berlin Heidelberg New York: Springer-Verlag, 1976.

[26] M. Wieling, "Analyzing dynamic phonetic data using generalized additive mixed modeling: a tutorial focusing on articulatory differences between l1 and l2 speakers of english," *Journal of Phonetics*, pp. 1–53, forthcoming.

[27] B. Haller, G. Schmidt, and K. Ulm, "Applying competing risks regression models: an overview." *Lifetime Data Analysis*, vol. 19, no. 1, pp. 33–58, 2013.