Interspeech 2018
2-6 September 2018, Hyderabad

Analyzing EEG signals in auditory speech comprehension using Temporal Response Functions and Generalized Additive Models

Kimberley Mulder[1], *Louis ten Bosch*[1], *Lou Boves*[1]

[1]Center for Language Studies, Radboud University, Netherlands
Kimberley.Mulder@let.ru.nl, l.tenbosch@let.ru.nl, l.boves@ru.nl

Abstract

Analyzing EEG signals recorded while participants are listening to continuous speech with the purpose of testing linguistic hypotheses is complicated by the fact that the signals simultaneously reflect exogenous acoustic excitation and endogenous linguistic processing. This makes it difficult to trace subtle differences that occur in mid-sentence position. We apply an analysis based on multivariate temporal response functions to uncover subtle mid-sentence effects. This approach is based on a per-stimulus estimate of the response of the neural system to speech input. Analyzing EEG signals predicted on the basis of the response functions might then bring to light condition-specific differences in the filtered signals. We validate this approach by means of an analysis of EEG signals recorded with isolated word stimuli. Then, we apply the validated method to the analysis of the responses to the same words in the middle of meaningful sentences.

Index Terms: ERP analyses, speech comprehension, reduced speech

1. Introduction

In psycholinguistics, event-related potentials (ERPs) are considered to be related to cognitive activities during the processing of spoken or written stimuli. For instance, the amplitude of the N400 ERP component is assumed to be inversely related to the cognitive effort required to semantically process the word to which the EEG is time-locked [1]. However, ERP signals not only reflect the brain activity related to semantic processing, but also multiple other activities, some of which may be related to other features of the stimuli (such as precontextual effects). A case in point is semantic processing of words embedded in continuous speech. It is highly unlikely that all confounding effects can be removed by conventional straightforward averaging over multiple tokens.

EEG signals are the response of an extremely complex – and probably non-linear– system to several simultaneous input signals, together with ongoing internal processes. Therefore, invoking knowledge from system identification theory should help in analyzing and understanding EEG signals, and the ERP responses derived from those signals. This idea was pioneered for studying responses of single cells as early as the nineteen eighties [2], and later adapted for non-intrusive investigation of whole-brain systems by [3, 4, 5] in Ireland and by [6, 7, 8] in the USA. By matching the physical characteristics of the stimuli with the features of the corresponding EEG signals it is possible to infer the system characteristics of a neural system. These characteristics are known as the Spectro-Temporal Receptive Field or Spectro-Temporal Response Function (STRF) or, in the parlance of the Irish group, the multivariate temporal response functions (mTRF). Recently, the two groups joined forces in an attempt to better understand the cocktail-party effect [9]. In this paper we investigate whether the STRFs can be used to separate concurrent neural activities from the same EEG signal, caused by auditory input that unfolds at the same time as the semantic processing takes place. The long-term goal of this research is to develop methods that will allow to conduct research on speech comprehension with stimuli that are much closer to everyday speech than the strictly controlled stimuli that are now being used in laboratory experiments. Because the mTRF software [5] is easier to access and use than the `STRFpak` and `STRFlab` packages designed by the group in the USA,[1] all processing was done using the mTRF software.

To understand the ways in which using the STRF concept can help to separate concurrent neural activities, we will first apply mTRF processing to EEG recordings of an experiment in which participants listened passively to multi-syllabic isolated words with a /ə/ in the first syllable that could be either reduced or full. Then, we will investigate EEG signals corresponding to the same words embedded in the middle of carrier sentences. We will combine mTRF processing with advanced statistical modeling using Generalized Additive Models (GAM), which are able to account for stimulus-related and participant-related variance [10, 11, 12].

2. Description of the data

The EEG data used for this paper was taken from [13, 14]. In three passive listening experiments, right-handed native listeners of Dutch were instructed to listen attentively to the presented speech input and were told that they would get questions about the words and sentences they were about to hear. The target words were full and reduced pronunciations of verb forms. The goal of the experiments was to test whether, and in which linguistic contexts, full forms have an advantage over reduced forms. That advantage was expected to show up in differences between two ERP components, an N100/P200 related to acoustic processing and an N400 related to semantic integration.

The full and reduced forms were presented in three listening contexts: the words presented in isolation, the words in mid-sentence and the words in sentence-final position. Participants took part in only one of the three experiments. Similar to [10], in this paper we only analyze full and reduced forms in isolation and in mid-sentence position.

2.1. Stimulus materials

The target stimuli were 80 Dutch verb forms starting with the unstressed prefixes be- (/bə/, e.g., bevallen /bəvɑlə/, to give birth), ge(/xə/, e.g., genieten /xəni:tə/, to enjoy), or ver- (/vər/, e.g., vertellen /vərtɛlə/, to tell). When pronounced in their full forms, these prefixes contain a clear ə. Only verb forms whose second syllable starts with a consonant were selected. Out of

[1] http://theunissen.berkeley.edu/

10.21437/Interspeech.2018-1676

the 80 verb forms, 31 had ver-, 31 be- and 18 had ge-. In addition, 120 filler verb forms were used that do not start with one of the three prefixes. In mid-sentence position, the verb forms served different syntactic functions to ensure that the results do not depend on a specific syntactic construction. The target verb form was always preceded by four syllables. Sentence accent was never on the target verb form or the preceding syllables. The semantic context up until the target verb form was kept as neutral as possible.

Sentences were recorded by a male native speaker of Dutch three times: Once without specific instructions, once with the instruction to pronounce all verb forms in full, and once with the instruction to pronounce the verb forms without the prefixal /ə/. For the filler sentences, there was no specific instruction. The reduced and unreduced verb forms were spliced out of their original sentences and were pasted into the carrier sentence or presented in isolation (these were segmented from the sentences in which these targets occurred at sentence-final position; see [13, 14] for more details). This was done to make sure that the reduced and unreduced sentences only differed with respect to the realization of the target verb form. The spliced, reduced and full verb forms had a mean /ə/ duration of 3 ms and 42 ms in mid-sentence, and of 0 ms and 43 ms in isolation, respectively. The mean duration of the words was 430 ms for the reduced forms and 495 ms for the full forms in mid-sentence position, and 739 ms and 782 ms in isolation, respectively.

2.2. EEG recordings

The EEG signals were recorded with 26 active electrodes mounted in an elastic cap (Acticap), two electrodes on the mastoids and four electrodes (two horizontally and two vertically placed) to capture the electro-oculogram (EOG). See [13, 14] for more details on the electrode montage. Each electrode was referenced online to the left mastoid. Electrode impedance was kept below 5 $k\Omega$. The EEG and EOG signals were amplified (pass band: 0.02 - 100 Hz), and digitized with a sampling frequency of 500 Hz. Before data analysis, the signals were re-referenced to the average of the left and right mastoids and digitally filtered with a low pass filter with cut-off at 30 Hz. Artifact detection and rejection was carried out using the criteria defined in the section "Raw Data Inspector" in the BrainVision User Manual [15]. The artifact rejection was applied to the channels of interest individually.

3. Temporal Response Functions

The mTRF approach described in [5] makes the simplifying assumption that the brain is a linear system that is completely identified by its impulse response, i.e., its response to the simplest possible input, which consists of a very brief excitation. This is reminiscent of the practice in EEG studies to use brief stimuli, with large refractory periods between stimuli. Impulses are so useful in system identification because the corresponding frequency spectrum is flat over a very wide bandwidth. White noise, with a sufficiently long duration, has the same advantage, and has also been widely used in system identification research. However, neither impulses nor white noise are useful for investigating the listeners' neural response to speech. Fortunately, it can be shown that the mathematics underlying system identification can be extended to arbitrary input signals. This makes it possible to obtain a useful estimate of the neural response to continuous speech [16].

Using the continuous EEG recordings from a complete stimulus, in parallel with the speech input, we compute an mTRF function for each individual stimulus. To synchronize the speech input with the EEG signals with 500 Hz sampling frequency we compute the loudness envelope of the speech, sampled at a rate of 500 Hz, and also low pass filtered at 30 Hz. For this purpose we used the software described in [17]. We use the mTRF functions to predict the EEG signals evoked by the full and reduced stimuli, without the non-stimulus related EEG activity. It is plausible to expect that differences between full and reduced stimuli, if they exist, will be more evident in the predicted output.

4. Analysis and results

We analyzed the EEG traces of the Cz electrode for the isolated and the mid-sentence stimuli by using Generalized Additive Modeling (GAM) [12, 18, 19]. Next to methods like growth curve analysis, functional data analysis and sparse functional linear mixed modeling, GAMs extend linear models in which a linear relationship between predictors and dependent variables is assumed. In a linear regression model all non-linear relations between the dependent variable and predictors must be specified by means of fixed algebraic expressions. In a GAM, a predictor can be expressed by a non-linear smooth function, and interactions can be combined in a multivariate (hyper)surface smooth. The advantage of GAMs over the manual specification of non-linearities in an `lmer()` is the flexibility: the optimal shape of the non-linearity is determined automatically, and the appropriate degree of smoothness can be determined on the basis of cross-validation to prevent overfitting. In GAMs, also random effects can be treated as smooths [12, 18].

To investigate whether in the mTRF data the full-reduced condition was more significantly different than in the non-mTRF data, we combined the raw (non-mTRF) and mTRF data and estimated a combined GAM model in which the the interaction between the two 2-level predictors ('without/with mTRF') and ('full/reduced') was included. The significance of this interaction provides information about whether the mTRF prediction significantly increases the difference between the full and reduced conditions. Models were always compared by using `compareML()`.

4.1. Isolated word data

For the isolated word stimuli (1,791,900 data points), the following GAM was found to be optimal:

$$\begin{aligned} GAMmodel = bam(amplitude\tilde{} \\ be.ge.ver + full.red * exp \\ +s(word\_dur, k = 20) + \\ s(t, by = full.red, k = 50) \\ +s(t, by = exp, k = 50) + ti(t, \text{Ə}\_dur) + \\ s(subject, bs = "re") + s(stimulus, bs = "re"), \\ data = data, samfrac = 0.1, gc.level = 2, \\ correlation = corAR1()) \end{aligned}$$

EEG amplitude serves as dependent variable, 'be.ge.ver' denotes the 3-level predictor denoting the prefix, 'full.red' denotes the predictor full/reduced, 'exp' denotes the categorical predictor without/with mTRF, t denotes the physical time (in 2 ms steps, from 200 ms before to 900 ms after onset), and 'ə_dur' denotes the duration of the /ə/. Subject and stimulus are used as random effects. Residual correlations are modeled away by

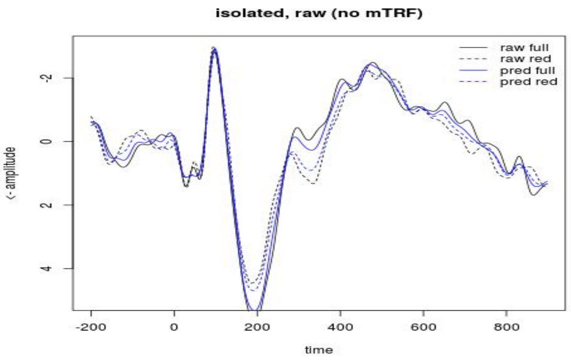

Figure 1: *Observed (black) and modeled (blue) EEG Cz trace for full and reduced isolated words, raw data. The horizontal axis displays time in ms; 0 refers to word onset. Dashed and solid lines refer to reduced and full stimuli.*

Table 1: *Parametric coefficients of the GAM modeled on isolated words.*

	Estimate	std. error	t	Pr(> $\|t\|$)
(Intercept)	-0.45126	0.31262	-1.444	0.148878
be.ge.ver [ge]	0.06531	0.24831	0.263	0.792526
be.ge.ver [ver]	0.45653	0.21054	2.168	0.030131
full.red [red]	0.07624	0.02290	3.329	0.000873
exp [raw]	0.42329	0.01708	24.778	< 2e-16
full.red [full] : exp [raw]	-0.11372	0.02414	-4.712	2.46e-06

Approximate significance of smooth terms

	edf	Ref.df	F	p
s(word_dur)	18.71	18.98	98.233	< 2e-16
s(t) : full.red [full]	11.22	13.73	0.566	0.931613
s(t) : full.red [red]	22.37	26.96	2.207	0.000311
s(t) : exp [mtrf]	26.78	31.93	3.408	2.63e-10
s(t) : exp [raw]	46.51	48.41	45.339	< 2e-16
ti(t, ə_dur)	15.59	15.97	33.115	< 2e-16
s(subject)	19.99	20.00	2548.440	< 2e-16
s(stimulus)	76.55	77.00	263.696	< 2e-16

the GAM function corAR(); the function `bam` is a fast implementation of a GAM [18]. The result on isolated word data is presented in Table 1.

Levels of factor predictors are shown between square brackets. This model explains 5.39% of the variance in the observed EEG data. It can be seen that not only the predictors 'without/with mTRF' and 'full/reduced' are significant, but in addition their interaction is significant. An analysis of the main predictor coefficients and the interaction in this table shows that the difference between full and reduced 'with-mTRF' is about twice as large as in the 'without-mTRF' condition.

4.2. Mid-sentence data

The finding that mTRF predictions improve the separation of full and reduced trials with isolated words justify the application of the same technique to the mid-sentence data, where previous analyses, based on raw EEG signals, did not uncover significant differences [13]. The mTRFs were computed on complete sentence stimuli; subsequently, the EEG signals for the complete sentence were predicted using the mTRF model, and 2000 ms intervals starting 200 ms before the onset of the target words were segmented from the predicted EEG signals.

Table 2: *Parametric coefficients, mid-sentence stimuli*

	Estimate	std. error	t	Pr(> $\|t\|$)
(Intercept)	-0.132	0.116	-1.132	0.258
be.ge.ver [ge]	-0.127	0.133	-0.954	0.340
be.ge.ver [ver]	0.242	0.013	18.733	<2e-16
full.red [red]	0.066	0.0112	5.606	2.07e-08

Approximate significance of smooth terms

	edf	Ref.df	F	p
s(word_dur)	18.96	19.00	262.68	<2e-16
s(time):full.red_fac0	28.15	34.22	22.18	<2e-16
s(time):full.red_fac1	26.21	32.05	13.36	<2e-16
s(subject_fac)	25.94	26.00	434.70	<2e-16
s(stimulus_fac)	77.80	78.00	417.94	<2e-16

A GAM model that simultaneously analyzed the mTRF predictions and the raw signals showed a significant interaction between the factors full.reduced and raw.mTRF. This significance justifies independent analyses of the raw and mTRF-predicted signals. In the model for the raw signals the factor full.reduced was not significant. The model for the mTRF predicted signals was defined as

$$\begin{aligned} GAMmidsentence = bam(amplitude\tilde{}\\ be.ge.ver + full.red\\ +s(word\_dur, k = 20)\\ +s(time, by = full.red, k = 50)\\ +s(subject, bs = "re")\\ s(stimulus, bs = "re"),\\ data = data, samfrac = 0.1, gc.level = 2,\\ correlation = corAR1()) \end{aligned}$$

As in [10] the GAM models were used to predict the mean Cz EEG traces. For the isolated word models, figure 1 shows the comparison between the mean observed raw (black lines) Cz EEG signal and the signal predicted by the GAM in Table 1 (blue lines). Figure 2 shows the mean observed and predicted mTRF-mapped EEG signals, modeled by the same GAM model. While the model is estimated on the combined datasets, it is able to follow the dynamic structure in the EEG quite well, both for full (solid lines) and reduced (dashed lines) stimuli.

5. Discussion and conclusion

The main goal of this paper was to investigate whether mTRF predictions of neural activity, which are supposed to offer a more focused view of the EEG activity induced by experimental stimuli, can uncover subtle differences between treatments when the conditions are adverse, i.e., when treatment-related stimulation overlaps with stimulation not related to the treatment. For the isolated word stimuli, where there is only little or no concurrent linguistic processing going on that is not related to the spoken words the GAM models already showed that the difference between full and reduced stimuli is enhanced by the mTRF predictions, relative to the raw EEG signals. In a previous study [10] we failed to find significant differences in the mid-sentence experiment. Here, we do find that mTRF-predicted EEG signals differ significantly between sentences with full and reduced versions of the target verbs. We attribute the difference to the details of the mTRF processing. In the previous study we analyzed pre-segmented epochs, for which we used the isolated word stimuli as the exogenous excitation. In addition, we estimated a single mTRF for all reduced and for all full stimuli. Here, we take into account the experience that participants in lengthy psycholinguistic experiments tend to vary their behavior and their attention quite substantially during the course of an experiment. That makes it reasonable to estimate mTRF responses on a local, stimulus-by-stimulus basis.

The differences between the full and reduced stimuli in the mid-sentence experiment are quite subtle (compare, e.g., Figs. 3, 4). In a sample-by-sample t-test we did not find time intervals in which the full and reduced stimuli were significantly different, neither in the raw, nor in the mTRF-predicted signals. The fact that a GAM model does show that the full and reduced conditions differ significantly shows that these differences are quite subtle, and only show up if additional confounding factors can be teared apart in the statistical analysis. Upfront, it was not clear whether there are differences in the mid-sentence condition, where both pronunciation variants of the verbs are appropriate (perhaps with a small preference for the reduced form). The fact that the tandem of mTRF prediction and GAM modeling uncovers differences opens interesting perspectives for future research with running speech.

In [10], sample-by-sample t-tests on the approximations by the GAM model showed that full and reduced forms differed significantly over different time intervals, but in the mid-sentence data only the interval around 150 ms after word onset was assumed to be potentially meaningful because it could be related to the P200 component assumed to reflect low-level, sensory/phonological processing of speech [20]. Upon visual inspection, the modeled EEG traces of the mTRF predicted signals fail to show this difference at 150 ms post word onset and they do not show a clear positive component in this time interval either. Possibly, the emergence of a P200 in [10] was an artifact, caused by the fact that the mTRF estimate was not based on complete sentences. Without guaranteed word segmentation in continuous speech the emergence of a P200 that is characteristic for isolated word processing may be quite unlikely.

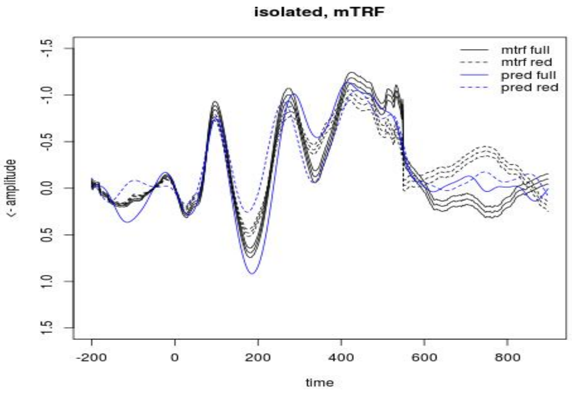

Figure 2: *Observed mean trace and predicted trace for full and reduced isolated words, mTRF data. Next to the mean, the deviation around the mean (1 sigma) is displayed for the observed data.*

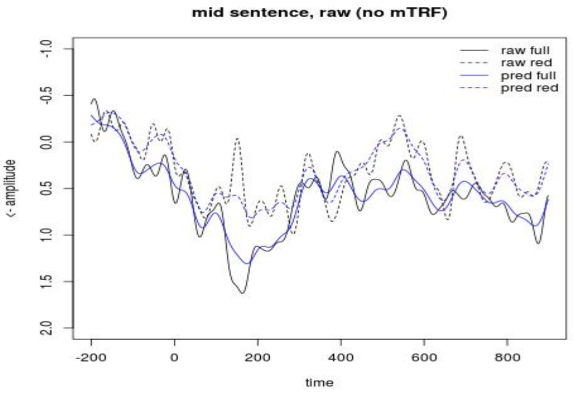

Figure 3: *Observed mean trace and modeled trace for full and reduced mid-sentence words, raw data (i.e., no mTRF).*

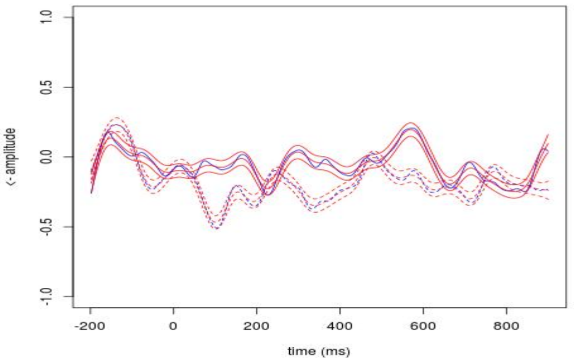

Figure 4: *Observed (blue) mean trace and modeled (red) trace for full (solid) and reduced (dashed) stimuli (plus 1σ bands, after mTRF). Predictions are obtained from the combined model.*

The mTRF prediction accounts only for a small proportion of the variance in the EEG signals. This is because stimulus/treatment-induced activity accounts for only a small proportion of the total EEG activity. Unsurprisingly, the proportion of the total variance that can be captured is larger in the isolated word data, where fewer nuisance factors are involved. However, both with isolated words and mid-sentence data, the coefficient of variation in the mTRF predictions is much smaller than in the raw signals. This suggests that it is interesting to pursue a more in-depth analysis of the contributions that spectro-temporal response functions can make to processing of EEG signals in psycho-linguistic research. Future plans include comparing participant-specific response functions.

It is also interesting to analyze additional information that is present in the GAMs. Preliminary analyses suggest that the smooths show systematic differences between listeners in the timing of effects in the EEG and the acoustic stimuli. These differences could be related to participant-specific mTRFs.

6. Acknowledgements

This work was supported in part by an ERC consolidator grant awarded to prof. Mirjam Ernestus (nr 284108).

7. References

[1] M. Kutas and S. A. Hillyard, "Brain potentials during reading reflect word expectancy and semantic association," 1984.

[2] J. Eggermont, P. Johannesma, and A. Aertsen, "Reverse-correlation methods in auditory research," *Quarterly Reviews of Biophysics*, vol. 16, pp. 341–414, 09 1983.

[3] E. Lalor, B. Pearlmutter, and J. Foxe, "Reverse correlation and the VESPA method," in *Brain Signal Analysis: Advances in Neuroelectric and Neuromagnetic Methods*, T. C. Handy, Ed. Cambridge, Mass.: MIT Press, 2009.

[4] N. Gonçalves, R. Whelan, J. Foxe, and E. Lalor, "Towards obtaining spatiotemporally precise responses to continuous sensory stimuli in humans: a general linear modeling approach to EEG," *NeuroImage*, vol. 97, pp. 196–205, 2014.

[5] M. Crosse, G. Di Liberto, A. Bednar, and E. Lalor, "The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli," *Frontiers in Human Neuroscience*, vol. 10, 2016, article 604.

[6] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 824–839, 1992.

[7] D. Klein, D. Depireux, J. Simon, and S. Shamma, "Robust spectrotemporal reverse correlation for the auditory system: Optimizing stimulus design," *Journal of Computational Neuroscience*, vol. 9, pp. 85–111, 07 2000.

[8] S. David, N. Mesgarani, and S. Shamma, "Estimating sparse spectro-temporal receptive fields with natural stimuli," *Network: Computation in Neural Systems*, vol. 18, no. 3, pp. 191–212, 2007.

[9] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *Cerebral Cortex*, vol. 25, no. 7, p. 1697, 2015. [Online]. Available: + http://dx.doi.org/10.1093/cercor/bht355

[10] K. Mulder, L. ten Bosch, and L. Boves, "Comparing different methods for analyzing erp signals," in *Proceedings of Interspeech*, San Francico, 2016.

[11] A. Tremblay and R. H. Baayen, "Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall," in *Perspectives on formulaic language: Acquisition and communication*, D. Wood, Ed. London: The Continuum International Publishing Group, 2010, pp. 151–173.

[12] S. N. Wood, *Generalized additive models*. New York: Chapman & Hall/CRC, 2006.

[13] L. Drijvers, K. Mulder, and M. Ernestus, "Alpha and gamma band oscillations index differential processing of acoustically reduced and full forms," *Brain and Language*, vol. 153-154, pp. 27–37, 2016.

[14] K. Mulder, L. Drijvers, and M. Ernestus, "The time course in processing reduced and unreduced word pronunciation variants: An ERP study," submitted.

[15] *BrainVision Analyzer User Manual, Version 1.03*, Brain Vision LLC, Morrisville, NC, 2006, page 125ff.

[16] E. Lalor and J. Foxe, "Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution," *European Journal of Neuroscience*, vol. 31, no. 1, pp. 189–193, 2010.

[17] B. Moore, "Development and current status of the 'cambridge' loudness models," *Trends in Hearing*, vol. 18, 2014.

[18] 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.

[19] S. N. Wood, *Generalized Additive Models: an introduction with R*. Boca Raton: CRC press, 2017.

[20] M. Rugg and M. Coles, *Electrophysiology of mind:Event-related brain potentials and cognition*. Oxford University Press, 1995.