Comparing different methods for analyzing ERP signals

Kimberley Mulder¹, Louis ten Bosch¹, Lou Boves¹

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

Abstract

Event-Related Potential (ERP) signals obtained from EEG recordings are widely used for studying cognitive processes in spoken language processing. The computation of ERPs involves averaging over multiple participants and multiple stimuli. Especially with speech stimuli, which also evoke substantial exogenous excitation, even averaging within conditions results in pooling many sources of variance. This raises questions about the statistical processing needed to uncover reliable differences between conditions. In this study we investigate differences between ERPs when participants listened to full and reduced pronunciations of verb forms in Dutch, in isolation and in mid-sentence position. Conventional statistical analysis uncovers some (but not all) differences between full and reduced forms in isolation, but not in mid-sentence position. In this paper, we show that linear mixed models (lmer) and generalized additive models (gam), which are able to account for participant- and stimulus-related variance may uncover more effects than conventional statistical models. However, depending on the complexity of the data, lmer and gam models may not be able to fit the data closely enough to warrant blind interpretation of the summary output. We discuss opportunities and threats of these approaches to analyzing ERP signals.

Index Terms: ERP based on EEG, reduced vs. unreduced speech, statistical analysis, generalized additive models

1. Introduction

Speech comprehension involves a complex sequence of processes for accessing information at different levels of representation. To study listening comprehension, one should be able to separate and account for the different cognitive processes that are involved during the unfolding of an auditory stimulus. Before the advent of sophisticated neurophysiological experimental techniques, studies on speech processing were mainly restricted to collecting button presses (i.e., that reflect response latencies or accuracy judgments) to isolated words or sentences. These experimental techniques provide output measures that are the sum of multiple cognitive (stimulus-related and stimulus-unrelated) processes that affect the processing of a stimulus. For instance, a response latency for a given stimulus not only includes the time it takes to actually process the stimulus, but also the decision about the response to be given, the motor planning and execution (pressing a specific button). This makes most behavioral experimental techniques unsuitable to study the processing of a stimulus proper, because they cannot separate the individual cognitive processes. Moreover, output measures do not capture the continuous nature of the speech signal and cannot account for the time course of effects on the processing of auditory stimuli (i.e., how different variables affect the processing of a stimulus as the signal unfolds).

One way for studying the time course of auditory processing is to use Event-Related Potentials (ERPs). ERPs provide a millisecond-to-millisecond record of the electrical changes evoked by an ongoing stimulus. With this technique, one can capture the continuous nature of the speech signal. However, ERPs are averages over multiple responses, almost always of multiple participants and/or multiple stimuli. Therefore, the averaging process is (perhaps tacitly) based on the assumption that the underlying processes are very similar in all participants and all stimuli. These assumptions justify the use of repeated measure ANOVA techniques for comparing average amplitudes of the ERP signals between experimental conditions. It is not unusual for those ANOVAs to fail to return significant differences between conditions that researchers expected to differ substantially. This has raised the question whether the assumptions underlying ANOVA do indeed hold, and whether more advanced analysis techniques can uncover effects that are not uncovered by ANOVAs. This paper addresses this question on the basis of an experiment aimed to investigate whether native listeners of Dutch process full and reduced verb forms differently.

2. Description of the experiment

Right-handed undergraduate students passively listened to full and reduced pronunciations of verb forms in three conditions: the words presented in isolation (N = 31, mean age = 21.3, sd age = 2.7), the words in mid-sentence (N = 28, mean age = 20.8, sd age = 2.0) and in sentence-final position (N = 27, mean age = 21.7, sd age = 1.8) [1, 2]. In this paper we only discuss the first two conditions. Stimuli were presented over Sennheiser HD215 headphones using the Presentation software (Neurobehavioural Systems, www.neurobs.com). Participants were instructed to attentively listen and were told that they would get questions about the words and sentences they were about to hear. The goal of the experiment was to test whether full forms have an advantage over reduced forms in all conditions. That advantage should show up in differences between two ERP components, an N100 related to acoustic processing and an N400 related to semantic integration.

2.1. Stimulus materials

We selected 80 Dutch verb forms starting with the unstressed prefixes be- (/be/, e.g., bevallen /bo’val/, ‘to give birth’), ge- (/ge/, e.g., genieten /x’a:nit/, ‘to enjoy’), or ver- (/✈/, e.g., vertellen /✈’til/, ‘to tell’). When pronounced in their full forms, these prefixes all contain a schwa. We only selected those verb forms whose second syllable start with a consonant. Out of the 80 verb forms, 31 with ver-, 31 with be- and 18 with ge-. In addition, we selected 120 filler verb forms in order to make the stimuli better represent the Dutch lexicon, which contains many verb forms that do not start with one of the three prefixes.

In mid-sentence position, the verb forms were the main
EOG signals were amplified (band pass = 0.10 - 100 Hz), and digitized with a sampling frequency of 500 Hz. Before data analysis, the signal was re-referenced to the average of the left and right mastoids and digitally filtered with a high cut-off filter of 30 Hz. Next, the continuous EEG was segmented into stimulus-time-locked epochs, starting from 200 ms before target word onset up to 800 ms after onset. Ocular artifacts were identified and removed with independent component analysis (ICA). After ICA, single trials that still contained artifacts were removed by a semi-automatic rejection routine. The period of 200 ms preceding the target word onset was used for baseline correction.

In this paper we confine ourselves to analyzing a single EEG signal, viz. Cz.

### 3. Isolated Words

Behavioral studies have shown that, at least in isolation, full forms benefit from a processing advantage over reduced forms (e.g. [3]). The ERP data from the isolated words for electrode Cz are shown in Figure 1. The green curve in the upper panel of Figure 1 shows the result of a sample-by-sample t-test on the average ERP signals from the full and reduced words. In a way, this corresponds to an ANOVA on very narrow bins, each containing a single sample (c.f. [4]). From the lower panel of Figure 1 it can be seen that there is a small processing advantage for full forms presented in isolation. More negative N100 amplitudes were observed for reduced forms (full red lines) relative to full forms (full blue lines). As speakers do not typically reduce words when uttered in isolation, the reduced forms could have been unexpected and, as a consequence, attracted more attention. A more negative N100 might also indicate speech segmentation problems [5], because absence of schwa in some reduced forms resulted in consonant clusters that only occur cross-word. However, as can be seen from the green curve in the upper panel of Figure 1, the amplitude difference between the ERPs for the full and reduced forms did not approach significance in the N100 region.

The green curve in the upper panel of Figure 1 suggests that the amplitudes of the ERP for the full and reduced forms differ significantly in the time interval from 300 to 400 ms. We interpret this as a difference in N400 peak latency for reduced and full forms, with a later N400 peak for reduced forms compared to full forms. As the N400 indexes lexical-semantic activation [6], this suggests that the activation of semantic representations is delayed for reduced forms. This is in line with behavioral findings that suggest that it takes more time to activate the semantic network for reduced forms [7].

For reduced forms, the N400 was preceded by a small negative peak around 270 ms, which was present at all central, fronto-central and centro-parietal sites and absent or less pronounced for full forms. We consider this peak to be the N250 [8], which might reflect a lexical selection process that occurs at the interface of lexical form and contextual meaning. Based on the incoming acoustic input, potentially matching lexical candidates are activated (e.g., [9]). [8] argue that an N250 effect would arise if contextual specifications do not support the activation of a lexical candidate, relative to a situation in which form-based activation is supported by the specifications of the context. The fact that words tend not to be reduced when they are pronounced in isolation therefore explains the finding that only reduced and not full forms generated a (clear) N250.

In sum, the ERP data seem to confirm a processing advantage for full forms presented in isolation. While a conventional ANOVA confirmed the expected difference related to the N400 component, it failed to uncover the – also expected – N100 component.

#### 3.1. Linear Mixed Effects Models

The failure to find a significant N100 effect may be due to several different causes. Maybe the most obvious cause is the true absence of an effect, which would suggest that full and reduced versions of the verbs heard as words spoken in isolation are equally likely and equally natural in normal speech. However,
we cannot exclude the possibility that there is some effect, but that it is obscured by pooling multiple sources of variance.

One way for investigating the contributions of several sources of variance is to replace ANOVA by a linear mixed effects model. For that purpose we used the R-package lme4 [10]. We started with a fairly simple model:

\[
m_0_{\text{lmer}} \leftarrow \text{lmer}(\text{amplitude} \sim \text{type}_{\text{fac}} + \text{type2}_{\text{fac}} + \text{word}_{\text{dur}} + \text{time} + (1|\text{subject}_{\text{fac}}) + (1|\text{stimulus}_{\text{fac}}), \text{data} = \text{dat})[-1\text{mm}]
\]

The output of that model is summarized in Table 1. The predictor \(\text{type}_{\text{fac}}\) corresponds to the planned difference between full and reduced forms; the predictor \(\text{type2}_{\text{fac}}\) has the three levels be-, ge-, ver-. We included \(\text{time}\) as a predictor, to prevent the model from collapsing all samples. We included \(\text{subject}\) and \(\text{stimulus}\) as random factors, because we aimed to remove as much participant- and stimulus-related variance as possible. From Table 1 it can be seen that the factor full/reduced is significant, as are the duration of the stimulus words and the factor \(\text{time}\). However, from the summary output it is not possible to know in what time interval the ERP amplitudes of the full and reduced forms differ significantly. We hoped to be able to glance that information from the predictions of the model of the amplitude of the ERPs of the individual stimuli. The averages of these predictions are shown by the dotted lines in the lower panel of Figure 1. However, these approximations turn out to be straight lines. Using hindsight, this does not come as a surprise: ERP signals are not linear functions of time, and as a result a strictly linear model cannot achieve close approximations.

An analysis without the factor \(\text{word}_{\text{dur}}\) returns only ver- and \(\text{time}\) as significant. We believe that this is due to the fact that the model must have at least basic information information about the temporal structure of the acoustic stimuli to be able to account for possible time shifts in the neural processes related to different stimulus durations.

### 3.1.1. Adding \(\text{time}\) as a non-linear predictor

The \(\text{lmer()}\) function in R offers the possibility of adding higher-order polynomial functions of predictor variables to the model. Because we have seen that ERP amplitude is not a linear function of time, we experimented with adding polynomials of time with increasing order. Table 2 shows a small part of the output of model \(m_2\_{\text{lmer}}\).

\[
m_2_{\text{lmer}} \leftarrow \text{lmer}(\text{amplitude} \sim \text{type}_{\text{fac}} + \text{word}_{\text{dur}} + \text{type}_{\text{fac}} \ast \text{poly}(\text{time}, 2) + (1 + \text{type}_{\text{fac}}|\text{subject}_{\text{fac}}) + (1|\text{stimulus}_{\text{fac}}), \text{data} = \text{dat})
\]

It can be seen that the factor \(\text{type}_{\text{fac}}\) is no longer significant in its own right. However, many of its interactions with polynomials of \(\text{time}\) are highly significant (not shown).

The dash-dot lines in the lower panel of Figure 1 show the approximation of the average ERP signals for full and reduced versions of the verb forms obtained with the model \(m_2\_{\text{lmer}}\). The black curve in the upper panel shows the \(p\) values from sample-by-sample \(t\)-tests for the difference between the approximated ERP signals. It can be seen that the model suggests that the amplitudes of the ERP signals pertaining to the full and reduced words differ significantly in several long, connected time intervals. However, from the lower panel of Figure 1 is it evident that the approximated ERP signals do not fit the raw signals very well. Maybe it should not come as a surprise that the approximations of the model do not reproduce the high-frequency oscillations in the raw averages; also, it is not clear whether these oscillations correspond to relevant cognitive processes. However, it is alarming that the approximation of the reduced ERP does not reproduce the N250 component, to which we did assign a cognitive interpretation. This raises questions about the validity of model \(m_2\_{\text{lmer}}\).

### 3.2. Generalized Additive Models

Adding polynomials of the predictor \(\text{time}\) to an \(\text{lmer()}\) model is a kludge, if only because predictions outside the time interval in the analysis will very likely be completely wrong. The recently introduced family of Generalized Additive Models (GAM) offers an attractive alternative, by replacing polynomials by regression splines (see [11] for an introduction to GAM analysis). GAM determines a linear and/or non-linear equation that strikes a balance between over-fitting and overgeneralizing a set of data through a process called penalized iteratively re-weighted least squares. The algorithm separates the parametric

---

**Figure 1:** Event-related potentials (ERPs) to full forms and reduced forms presented in isolation for electrode Cz. Negativity is plotted upwards. Top panel: by-sample \(t\)-tests. X-axis represents time in ms. For details, see text.

**Table 1:** Summary of the \(m_0\_{\text{lmer}}\) model.

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>(t) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-1.6022229</td>
<td>0.616846</td>
</tr>
<tr>
<td>type_fac2</td>
<td>0.5055082</td>
<td>0.0508597</td>
</tr>
<tr>
<td>type2_fac2</td>
<td>-0.0402940</td>
<td>0.4675198</td>
</tr>
<tr>
<td>type2_fac3</td>
<td>0.9116503</td>
<td>0.3964945</td>
</tr>
<tr>
<td>word_dur</td>
<td>0.0079978</td>
<td>0.0005186</td>
</tr>
<tr>
<td>time</td>
<td>-0.0116187</td>
<td>0.0001587</td>
</tr>
</tbody>
</table>

**Table 2:** Partial summary of the \(m_2\_{\text{lmer}}\) model.

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>(t) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-5.335e+00</td>
<td>6.427e-01</td>
</tr>
<tr>
<td>type2_fac2</td>
<td>4.663e-02</td>
<td>4.762e-01</td>
</tr>
<tr>
<td>type2_fac3</td>
<td>9.478e-01</td>
<td>4.039e-01</td>
</tr>
<tr>
<td>word_dur</td>
<td>8.363e-03</td>
<td>5.110e-04</td>
</tr>
<tr>
<td>type2_fac</td>
<td>5.835e-01</td>
<td>3.797e-01</td>
</tr>
</tbody>
</table>
from the nonparametric part of the fit, and fits the parametric part using weighted linear least squares.

[12, 13] applied GAMs to analyze EEG signals, using the R-package mgcv [14]. We used the same software to obtain the GAM model \(\text{gam2} \).

\[
\text{gam2} \leftarrow \text{bam}(\text{amplitude} \sim \text{type} \cdot \text{fac} + \text{type} \cdot 2 \cdot \text{fac} \\
+ \text{s}([\text{word} \cdot \text{ur}] + \text{s}([\text{time}, \text{by} = \text{type} \cdot \text{fac}], k = 50) \\
+ \text{f}([\text{time}, \text{schwa} \cdot \text{ur}], \text{bs} = \text{"re"}) \\
+ \text{s}([\text{stimulus} \cdot \text{fac}, \text{bs} = \text{"re"}], \text{data} = \text{dat})), \\
\text{samfrac} = 0.1, \text{gc.level} = 2, \text{correlation} = \text{corAR1}())
\]

The approximations of the raw ERP signals obtained with this model are shown as dashed lines in the lower panel of Figure 1. If the dashed lines are extremely difficult to distinguish from the full lines, that is because the approximation with \(\text{gam2} \) is close to perfect. That is not to say that the approximations of individual EEG signals is very good. \(\text{gam2} \) only accounts for slightly less than 10% of the total variance in the data.

The \(p\)-values of a sample-by-sample \(t\)-test on the approximated signals again shows substantial connected time intervals in which the ERP signals of the full and reduced words, with at least part of the participant- and item-related variance removed, differ significantly. Specifically, in the GAM analysis the difference in the N100 region appears to be significant. Because of the close approximation of the average ERP signals by the GAM model we trust that here the \(p\)-values are meaningful.

4. Mid-sentence data

In [2] it was found that in mid-sentence position conventional ANOVAs did not discover differences between the ERP signals of full and reduced forms. This finding was explained by the fact that in mid-sentence position reduced pronunciations are at least as natural as full forms. However, it is worthwhile investigating whether significant differences can be found if part of the subject- and item-related variance removed, differ significantly. Specifically, in the GAM analysis the difference in the N100 region appears to be significant. Because of the close approximation of the average ERP signals by the GAM model we trust that here the \(p\)-values are meaningful.

The sample-by-sample \(t\)-tests on the approximations by the \(\text{lmer()} \) and \(\text{gam} \) models are shown in Figure 2. The full green line in the top panel confirms that \(t\)-tests do not detect significant differences in the raw ERPs.

The sample-by-sample \(t\)-tests on the approximations by the \(\text{lmer()} \) and \(\text{gam} \) models suggest that the full and reduced ERP signals differ significantly over several connected time intervals. However, from the lower panel in Figure 2 it is evident that both model approximations differ considerably from the raw averages of the full and reduced ERP signals. It can be argued that these averages are so complex (have so many local minima and maxima) that both models fail to capture the structure if any in the data. The adjusted \(R^2 \) of the \(\text{gam2} \) model applied to the mid-sentence is only 4.6%. It seems that the EEG signals corresponding to the words in mid-sentence position contain a substantial amount of variance that is not related to the predictors used in the model. A probable source is the exogenous excitation due to the speech preceding the interval under analysis.

Omitting the lessons learned from the analyzes of the isolated word data, we refrain from interpreting the differences that are significant according to the \(\text{lmer()} \) and \(\text{gam} \) models. Only the interval around 150 ms after the onset of the crucial word, where the analysis of the raw signals approaches significance, is likely to be meaningful. The larger negative-going amplitude of the full forms suggests that in mid-sentence position reduced forms are more likely.

In [15] it is recommended to evaluate the quality of \(\text{gam} \) models by analyzing the autocorrelation in the prediction error. Although this may work well when predicting time series of reaction times, which are indeed characterized by a first-order auto-regressive process that should and can be accounted for by a linear mixed effects model [16, 17], the situation with EEG signals is very different. Figures 1 and 2 show that the average ERP signals related to full and reduced pronunciations are much more complex than a first-order auto-regressive process. Especially for the mid-sentence data there is a need for additional processing to remove the exogenous excitations. The autocorrelation that is naturally present in EEG signals can only be removed by a model that accurately predicts all individual signals. No linear model with the predictors that we have available can be expected to accomplish this. We plan to investigate whether the linear deconvolution approach proposed in [18, 19] can remove enough exogenous excitation from the EEG signals to yield average ERP signals that come closer to a sequence of established ERP components.

5. Conclusions

In this paper we investigated several statistical procedures for analyzing EEG/ERP signals. We want to use such models to reduce the subject- and item-dependent variance sources that may obscure true differences between conditions, such as listening to full and reduced pronunciations of words. We found that linear models can only begin to approximate EEG signals if time can be included as a non-linear predictor. In \(\text{lmer} \) models this can be accomplished by introducing \(\text{poly}(\text{time}, n) \), where \(n \) specifies the order of the polynomial. Although such \(\text{lmer} \) models suggested that there are substantial connected time intervals where full and reduced pronunciations differ, we concluded that the fit of the model to the data was not good enough to be able to interpret the results. For the isolated words the \(\text{gam} \) model did obtain a very close approximation. However, the mid-sentence signals appear to contain so much variation that is not related to the predictors that even a gam model fit is not good enough.

6. Acknowledgements

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


