**Interactions of speech and speaker processing in early evoked components**

Paula Rinke¹,³, Nadine Lavan² and Mathias Scharinger¹,³

¹Research Group »Phonetics«, Institute of German Linguistics, Philipps-Universität Marburg, Germany
²School of Biological and Behavioural Sciences, Queen Mary University of London, United Kingdom
³Center for Mind, Brain & Behaviour »CMBB«, Universities of Gießen & Marburg, Germany

paula.rinke@uni-marburg.de, n.lavan@qmul.ac.uk, mathias.scharinger@staff.uni-marburg.de

**ABSTRACT**

The voice of a speaker and the linguistic content of a message are linked to one another during speech production and perception, allowing to dissociate between what is being said by whom. The time course of the interaction between speech and speaker information is, however, less well investigated. Using time-sensitive neuro-physiological measurements (EEG) with Event-Related Potentials (ERP), the present study aims to address this question.

16 participants attended a 1-back-task while being presented with recordings of the German vowels [a], [i], [u] from 32 German-speaking men and women, falling into a younger or older age group, (96 stimuli in total), while brain activity was recorded from 32 active electrodes.

ERP-analyses showed that vowel category elicited strong effects in early evoked components, especially in the N1 time window (100ms after stimulus onset), with speaker age and gender exerting significant modulations. This suggests an early interaction of speech and speaker processing.

**Keywords**: speaker processing, voice processing, EEG, ERP

1. **INTRODUCTION AND BACKGROUND**

Speech and speaker information interact with one another. Even though the voice is often simply considered as the “carrier of speech” [1] its influence on speech processing is inevitable. The voice carries much social information (so-called “indexical information”) about the speaker, e.g., gender, age, emotional state and social or regional background. However, the perception of speaker-specific features is often considered as higher-order processing, that is separate from low-level speech processing.

Previous studies, however, show that speaker-specific cues can affect low-level speech processing such as speech sound categorization [2]. For example, Strand [3] provides evidence that the gender attributed to a speaker influences the categorization of English [s] and [ʃ]. Moreover, Drager [4] shows that the perception of New Zealand English vowels is influenced by higher-order knowledge such as the perceived age.

These results provide evidence that the perceived social characteristics of a speaker affect speech processing. Thus, speech and speaker perception appear as interdependent rather than two completely separated processes.

Aside from these behavioral data, evidence on the neural basis of this interaction is comparably small. Speech information, such as vowel category, are known to be retrieved during very early processing stages. As previous neurophysiological studies (electroencephalography (EEG) and Event-Related Potentials (ERP) as well as magnetoencephalography (MEG)) have shown, spectral information inherent in different vowel categories (e.g., formant frequencies) affect early brain responses such as the N1 [5, 6], an auditory evoked component with a negative peak between 80 and 150 ms after stimulus onset [5]. In particular, the first formant (resonance frequency of the vocal tract), F1, was found to modulate amplitude and latency of the N1 [7].

While the time course of speech processing is well investigated, research on the time course of social characteristics of the voice is relatively rare. The few existing EEG studies on voice processing found two neural components reflecting the processing of speaker-specific features: the Mismatch Negativity (MMN) and the P3a [8-10]. The MMN is an early and automatic brain response with a negative deflection that is sensitive to change detection, whereas the P3a is a later component with a positive peak approximately 300 ms after stimulus onset and is considered to indicate involuntary attention switches caused by a novel stimulus [11].

In the current study, we brought together the investigation of speech and speaker perception using time-sensitive EEG and ERP measurements. By manipulating speech-related information (vowel categories) and speaker information (age and gender), we attempted to characterize the time course of
speech and speaker-related information. Based on the findings of previous ERP studies, we would expect vowel processing to be reflected in early brain components, such as the N1, with speaker’s age and gender possibly modulating the response pattern.

2. METHOD

2.1 Stimuli

96 recordings of the German vowels [a], [i], and [u] from 32 speakers (16 male; 16 female) were selected from the “Saarbrücker Stimmdatenbank” (http://www.stimmdatenbank.coli.uni-saarland.de/). Each group was equally divided into two age groups representing either younger (20-34 years) or older adults (55-74 years; details are listed in Tab. 1). Consequently, each category (e.g., male-old or female-young) consisted of eight different speakers with each one producing the three vowels (in total 24 sounds per vowel category).

<table>
<thead>
<tr>
<th>Category</th>
<th>Mean</th>
<th>Range</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female young</td>
<td>24.13</td>
<td>20-34</td>
<td>4.86</td>
</tr>
<tr>
<td>Male young</td>
<td>24.25</td>
<td>21-28</td>
<td>4.35</td>
</tr>
<tr>
<td>Female old</td>
<td>62.63</td>
<td>55-74</td>
<td>5.66</td>
</tr>
<tr>
<td>Male old</td>
<td>63.00</td>
<td>56-72</td>
<td>4.80</td>
</tr>
</tbody>
</table>

Table 1: Overview of different age groups in stimulus material

The German vowels [a], [i], and [u] are distinct in their tongue position, openness and backness with each vowel falling into easily separable locations in the German vowel space. They thus show distinctive spectral properties such as systematic differences in vowel formants, especially in F1 and F2.

Sounds were edited in Praat [12] and adjusted to an intensity level of 75 dB. Furthermore, all sounds were trimmed to a duration of 400ms in Audacity.

2.2 Experimental Design

Participants were presented with all 96 vowel stimuli, which were repeated 40 times across six blocks. During stimulus presentation, participants were asked to complete a 1-back-task for speaker identity. To test for continuous attention, we included vigilance trials for which additional vowel stimuli were selected (24 altogether). Vigilance stimuli were repeated 16 times and evenly distributed over the six blocks. In total, participants were presented with 4224 vowel stimuli.

Participants were placed in a shielded room approximately 1 m in front of a screen on which the instructions and the task were presented. The 1-back-task was completed via press of keyboard keys to indicate whether the presented and the preceding stimuli were the same (same vowel and speaker). Sounds were delivered via loudspeakers placed on the right and left side of the screen. To prevent extensive eye movement, participants were asked to focus on a white fixation cross on the screen during stimulus presentation and try to blink only right after responding to the task.

The experiment lasted approximately 70 minutes.

2.3 Participants

16 participants took part in the study (8 males, 8 females; mean age 26.13 years, range 19–36; standard deviation [SD] 5.12 years). All participants were native German speakers and had no reported history of hearing or neurological disabilities. Written informed consent was collected from all participants in order to participate in the study. The study was approved by the local ethics committee.

2.4 EEG recording

Continuous EEG was obtained from 32 active electrodes placed on a standardized 10–20 system cap. The reference electrode was placed on the nose while the ground electrode was placed on the forehead between Fp1 and Fp2 positions. Impedance was kept below 10 kΩ while EEG signals were sampled at 1000 Hz and online pre-filtered between 0.016 and 250 Hz using the Brain Vision Recorder software (Brain Products, Gilching, Germany), running on a Windows 10 personal computer.

2.5 EEG preprocessing

Continuous EEG was preprocessed using fieldtrip [13]. First, erroneous EEG channels (maximally 2 per participant) were identified and interpolated. Next, eye movements and blinks were identified through independent component analysis (ICA). Again, maximally 2 eye components per participant were identified and removed. Subsequently, the continuous EEG was split in epochs aligned to vowel onsets, starting 100 ms before and ending 700 ms after vowel onset. Epochs were bandpass-filtered (between 0.01 and 30 Hz) and the mean voltage of the 100-ms baseline was subtracted. Epochs were then averaged across repetitions within each participant and thereby corresponded to ERPs in a 3 (vowels [a], [i], [u]) x 2 (speaker age: old, young) x 2 (speaker sex: male, female) setup.
2.6 Statistical analyses

Visual inspection of the preprocessed data revealed two substantial time windows between 90 and 150 ms and 350 and 500 ms that correspond to the time windows of the N1 and the N4/P3. The statistical analysis was carried out for these two time windows and was based on the mean values of each item (i.e., vowel per speaker and age group). Vigilance trials were excluded; thus, the analysis was done for test items only. Mean amplitude in the time windows of interest and the latency of the highest N1 peak were calculated.

Linear mixed models (LMMs) were based on the Fz electrode and calculated for N1 amplitude and latency and for P3/N4 amplitude with vowel category, speaker’s gender and age as fixed effects and subject as random intercept. Data were analyzed in Jamovi [14].

2.7 Results

2.7.1 Amplitude

The LMM on N1 amplitude (Tab. 2) showed a main effect for gender (p<.001) and vowel category (p<.001), but not for age (p>0.2). N1 amplitudes were more negative for male than for female voices, and more negative for [i] than for the other two vowels. The interaction of gender and vowel category (p=.018) was driven by larger between-vowel differences for female than for male voices.

The interaction of gender and age (p=.014) showed that the gender effect was more pronounced for the young compared to the old speakers. No other interactions were significant (all p-values>0.2).

2.7.2 Latency

For N1 latency (Tab. 4), there were significant main effects of gender (p=.021) and vowel category (p<.001), but no age effect (p>0.2). Notably, the vowel [a] elicited earlier N1 responses than [i] or [u]. Vowels pronounced by female speakers also elicited earlier N1 responses than vowels pronounced by male speakers. No interactions were significant (all p-values>0.1).
As a consequence, [a] and [i] also show differences in front vowel but rather retracted as central vowel [u]. However, note that German /a/ is not realized as a front vowel but rather retracted as central vowel [a]. As a consequence, [a] and [i] also show differences in F1 and F2 properties. Thus, the current results cannot be exclusively attributed to one particular formant frequency.

The speaker’s gender was found to be relevant for both observed time windows and in particular for the amplitude and latency of the N1. Vowels produced by a male voice elicited a stronger but later N1 than vowels produced by female voices.

Since a person’s gender is often considered as “most primary judgments that human perceivers make of each other” [3] it is not surprising that its influence stretches across multiple time windows.

From an acoustic point of view, gender is strongly (but not solely) associated with the fundamental frequency (F0). F0 plays an important role in vowel perception besides fundamental frequencies. This is also displayed in the present study by the strong interaction of gender and vowel category for N1 amplitude. Thus, it is plausible that the speaker’s gender is retrieved during early speech processing reflected by the N1 as well as during speech processing displayed by the N4.

When focusing on the perceived age of the speaker, no effects on the amplitude and latency of the early brain response N1 as well as the later N4 can be observed. Thus, as expected from previous voice studies, the speaker’s age does not show any influence on early brain components which are mainly modulated by vowel categories and gender. However, significant interactions with the speaker’s gender were found for N1 amplitude, with a stronger effect of gender for younger than older speakers, which again highlights the importance of gender for early processing stages.

Regarding the interpretation of the gender effect on N1 properties, it is important to point out a possible limitation of these analyses. The N1 is an early component that is sensitive to acoustic changes. Since male and female voices differ substantially in F0 and other acoustic characteristics [15], we note that the differences found in amplitude and latency could at least partially reflect these low-level acoustic differences.

However, there were additional effects of gender in the later N4 time window. This time window is associated with higher-level processing (as opposed to early acoustic processing), which we cautiously interpret as a result of the processing of the speaker characteristic “gender”.

Overall, the current data confirm that speaker characteristics can in some cases influence early and later brain responses to speech-related information. The data thus show that speaker and speech-related information in fact routinely interact and inform each other during the perception of human vocal stimuli.
4. REFERENCES


ACKNOWLEDGMENTS

The research for this paper was funded in part by the German Research Foundation (Research Training Group 2700: »Dynamics and stability of linguistic representations«) and a Sir Henry Wellcome Fellowship (220448/Z/20/Z) awarded to Nadine Lavan.