Speaker Change Detection Based on Event-Related Potentials with a Consumer Brain-Computer Interface

Daniel Neudek[1], Anil Nagathil[1], Stephan Getzmann[2], Rainer Martin[1]

[1] *Ruhr-Universität Bochum, Institute of Communication Acoustics, Bochum, Email: firstname.lastname@rub.de*

[2] *Leibniz Research Centre for Working Environment and Human Factors - IfADo, Dortmund, Email: getzmann@ifado.de*

Introduction

The automatic detection of speaker changes in cocktail party scenarios can be beneficial for adapting speech processing algorithms in hearing devices or hearables. Different approaches for speaker change detection in the audio signal domain have been developed so far. For example, solutions based on mel-frequency cepstrum coefficients in combination with Gaussian mixture models [1, 2] or based on neural networks [3] were proposed. However, in more adverse acoustic conditions such algorithms need additional information to achieve a robust performance.

Additional information could be extracted from electroencephalography (EEG) recordings and in particular from event-related potentials (ERP). An ERP can be characterized by the amplitudes and latencies of positive and negative deflections obtained after stimulus onset. In a typical experimental oddball paradigm, a sequence of regular (standard) stimuli is interrupted by irregular (deviant) stimuli. In case of a speaker change, a deviant speaker interrupts the current (standard) speaker. Computing the difference waveforms of deviant and standard stimuli results in an ERP representation with deflections in specific time intervals after stimulus onset, which typically consists of the mismatch negativity (MMN) and P3a. The MMN is a negative deflection in the ERP at about 200 ms after stimulus onset [4]. It is most pronounced over frontal brain areas and is elicited by acoustic stimuli changes, when the current stimulus does not match the expected stimulus [3, 5, 6, 7]. The frontal component of the P3 peak (P3a) is a positive deflection in the ERP and occurs at about 300 - 500 ms after stimulus onset. A P3a peak is elicited, when a rare, non-predictable stimulus change occurs and the participant focuses his/her attention on the new stimulus [8]. Due to neural noise, the signal-to-noise ratio (SNR) of single-trial ERPs is typically low, so that ERPs are usually averaged over several trials in which a specific stimulus was presented [8].

In neuroscientific research, EEG signals are typically recorded with an EEG cap, using a standardized electrode system. Further, a high contact quality is achieved by minimizing the resistance between electrode and scalp. For real-world applications, brain-computer interfaces (BCI) with fewer electrodes and wireless connectivity have been developed to access the EEG signals and translate them into computer commands. The purpose of a BCI is to achieve a robust recognition of commands based on noisy, single-trial EEG signals, so that methods for signal quality enhancement become important [9].

In this work we investigated if speaker changes can be detected based on single-trial ERP signals. To this end, we used a consumer BCI (Emotiv Epoc+ [10]) to record EEG signals in an oddball paradigm with male and female speakers. After performing a pre-processing step, features for the characterization of the MMN and P3a were extracted and a linear discriminant analysis based classifier was trained to discriminate between standard and deviant speakers.

EEG Experiment

In total, $N = 10$ participants (8 male, 2 female) with an average age of 25.7 ± 4.14 years participated in this study. All participants were right-handed, healthy and had self-reported normal hearing. The experiment was approved by the Ethics Committee of Ruhr-Universtät Bochum (registration number 18-6376). Written informed consent was obtained before study commencement. We used an active oddball paradigm for the speaker change experiment with the word *two* spoken by two male and two female speakers. A continuous sequence of the word *two* was presented by either a regular standard speaker or by rare deviant speakers. The standard speaker changed from participant to participant and was presented in 85% of the trials, while the deviant speakers (remaining speakers) were played back in 15% of the trials. The participants were instructed to count each deviant stimulus. In total, we presented 720 stimuli with a 1s stimulus-onset asynchrony.

We performed the experiments in an audiometric booth with reduced acoustic reflections and electro-magnetic shielding. A MATLAB [11] script played back the acoustic stimuli in mono through a loudspeaker (GENELEC 2029b [12]), which was positioned in front of the participant and above the screen. For the EEG acquisition, we used the Emotiv EPOC+ device (channels AF3, AF4, F7, F3, F4, F8, T7, T8, P7, P8, O1, O2 and mastoid as reference, 256Hz sampling rate) with the software EmotivPro [10], while the stimuli were marked by a virtual serial interface from MATLAB. Note, that in [7] the previous model (Emotiv EPOC) was also used for recording ERP signals.

To illustrate the performance of the Emotiv EPOC+ device, we computed the waveform averages across all standard and deviant stimuli, respectively. Subtracting the averaged standard waveforms from the averaged deviant waveforms yielded a difference waveform. Finally, a grand average (GA) waveform was obtained by averaging across all participants. Figure 1 shows this GA waveform for the frontal electrodes F3 and F4, which al-

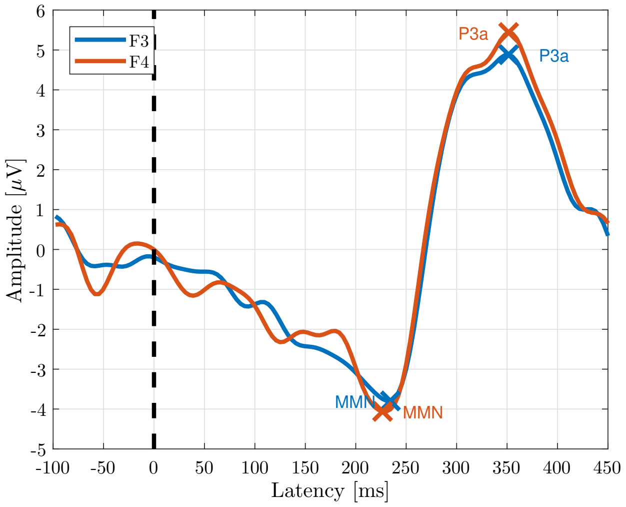

Figure 1: Difference grand-average waveform (deviant minus standard trials) for the channels F3 and F4. A Butterworth band-pass filter was used with $f_{\text{stop1}} = 0.1\text{Hz}$, $f_{\text{pass1}} = 1\text{Hz}$, $f_{\text{pass2}} = 25\text{Hz}$, $f_{\text{stop2}} = 30\text{Hz}$, $A_{\text{stop1}} = 60\text{dB}$, $A_{\text{pass}} = 1\text{dB}$, $A_{\text{stop2}} = 80\text{dB}$. The MMN and P3a occur approx. 230 ms and 350 ms after stimulus onset, respectively.

lows to determine the MMN and the P3a as negative and positive deflections peaking 230ms and 350ms after stimulus onset, respectively. The shown GA ERP indicates that there is a difference in the ERP for standard and deviant stimulus. This clear evidence results from averaging out the uncorrelated neural noise in multiple trials, which leads to a higher signal-to-noise ratio (SNR) [8].

Single-trial detection of speaker change

We can also assume systematic differences between single-trial ERPs obtained from a standard and a deviant stimulus, which however are more obfuscated in the noisier ERP signals. To detect such differences, we performed a pre-processing step and then extracted features from the enhanced ERP signals.

In the pre-processing step, we first filtered the raw EEG data with a Butterworth high-pass ($f_{\text{stop}} = 0.1\text{Hz}$, $f_{\text{pass}} = 1\text{Hz}$, $A_{\text{stop}} = 80\text{dB}$) and segmented then the single-trial ERPs with respect to the stimulus onsets from the EEG signal. Because the standard stimulus was presented most of the time, we applied a first-order recursive filter on the succession of the single-trial ERPs to obtain an enhanced version for the waveform of the standard stimuli. In the next step, we averaged the channels F3 and F4 for the long-term standard ERP and the single-trial deviant ERP, as MMN and P3a are most prominent at the frontal electrodes. Then we computed the deviation from the current single-trial ERP and the long-term averaged standard ERP of the last segment.

After performing the pre-processing step we extracted four signal features by computing the means and variances within time ranges surrounding the expected MMN and P3a peaks. From Figure 1, we considered the time ranges [150ms, 300ms] and [250ms, 400ms], respectively. We could not use the peaks and latencies for the MMN and the P3a of the single-trial ERPs, because the signal was too noisy. The first and second order statistics provides more robust features.

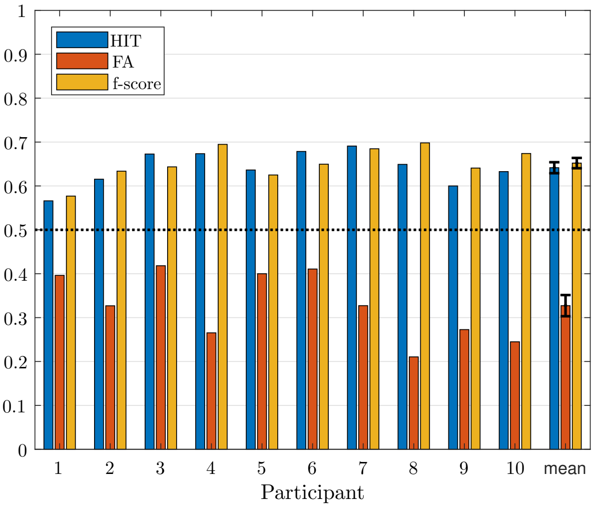

Figure 2: Results for the speaker change detection algorithm for each participant individually and as mean over all participants. Error bars in the averaged results denote the standard error of the mean.

As a classifier, we used the linear discriminant analysis for each participant individually and divided our stimuli into a training and a testing set with approx. 50 stimuli of each class and set. To obtain a balanced number of standard and deviant stimuli, we only considered the standard stimuli right before the deviants. As evaluation measures we computed the HIT rate (correctly detected deviants), the false alarm (FA) rate (standards miss-classified as deviant) and the f-score.

Results

Figure 2 shows the results of the speaker change detection algorithm based on the single-trial ERPs. Due to our design, the *a priori* probability for both classes and therefore the chance level for HIT and FA is 0.5. The results indicate that for all participants the algorithm has a similarly good performance (HIT>0.5 and FA<0.5). On average, both the HIT rate and FA rate significantly deviate from chance level ($p < 0.01$, obtained by t-test).

Conclusion

The results of our proposed detection algorithm show that we achieve a fairly robust level of performance across all participants. On average speaker changes can be detected significantly better than chance using a BCI and single-trial ERPs. This additional information potentially allows speech processing algorithms to adapt better to dynamic acoustic scenes in a cocktail party scenario. A remaining challenge resides in increasing the SNR, which will be tackled by developing improved noise reduction techniques. Furthermore, future work should consider continuously spoken speech to approach a more realistic scenario. In the long run, we believe that EEG-assisted speech processing algorithms can be beneficial for improving speech intelligibility and reducing listening effort in future hearing devices and hearables.

References

[1] Vuorinen, O., Peltola, J., and Makela, S.: Unsupervised Speaker Change Detection for Mobile Device Recorded Speech. IEEE International Conference on Acoustics, Speech and Signal Processing (2007), II-757-II-760

[2] Yutai, W., et al.: Speaker recognition based on dynamic MFCC parameters, International Conference on Image Analysis and Signal Processing (2009), 406-409

[3] Hrúz, M., Zajíc, Z.: Convolutional Neural Network for speaker change detection in telephone speaker diarization system. IEEE International Conference on Acoustics, Speech and Signal Processing (2017), 4945-4949

[4] Titova, N., Näätänen, R.: Preattentive voice discrimination by the human brain as indexed by the mismatch negativity. Neuroscience Letters 308 (2001), 63-65

[5] Näätänen, R., et al.: The mismatch negativity (MMN) in basic research of central auditory processing: A review. Clinical Neurophysiology 118 (2007), 2544-2590

[6] Berti, S., Roeber, U., Schröger, E.: Bottom-Up Influences on Working Memory: Behavioral and Electrophysiological Distraction Varies with Distractor Strength. Experimental Psychology 51 (2004), 249-257

[7] Badcock, N., et al.: Validation of the Emotiv EPOC® EEG gaming system for measuring research quality auditory ERPs. PeerJ 1:e38 (2013)

[8] Luck, S.: An Introduction to the Event-Related Potential Technique. MIT Press, 2014

[9] Selim, A., Wahed, M., Kadah, Y.: Electrode reduction using ICA and PCA in P300 Visual Speller Brain-Computer Interface system. 2nd Middle East Conference on Biomedical Engineering (2014), 357-360

[10] Emotiv homepage, URL: `https://www.emotiv.com/`, last accessed on 23 March 2020

[11] MATLAB product page, URL: `https://de.mathworks.com/products/matlab.html`, last accessed on 23 March 2020

[12] GENELEC homepage, URL: `https://www.genelec.com/`, last accessed on 23 March 2020