Neural Speech Processing During Selective Listening in an Audio-Visual Monologue vs. Dialogue Paradigm

AUTHORS: Sascha Bilert[1,2], Jens Hjortkjær[2], Emina Aličković[1,3], Martha Shiell[1], Sergi Rotger-Griful[1], Johannes Zaar[1,2]
[1]Eriksholm Research Centre, [2]Technical University of Denmark, [3]Linköping University

Multiple studies have shown that it is possible to decode an attended speech stream from single-trial electroencephalography (EEG) recordings (e.g. O'Sullivan et al. 2015). Previous research has mainly focused on controlled conditions often focusing on one out of two competing talkers. In the present study, we consider attention decoding in a more realistic condition by introducing an audio-visual multi-talker target (i.e. dialogue) versus a competing single talker.

Motivation

- A monologue (mono) vs. dialogue (dia) paradigm introduces a more realistic condition
- Multi-talker babble background noise creates a cocktail party like situation (e.g. Das, Bertrand, & Francart, 2018)
- Presenting audio-visual stimulation in this setup provides a multimodal condition introducing eye-movements.

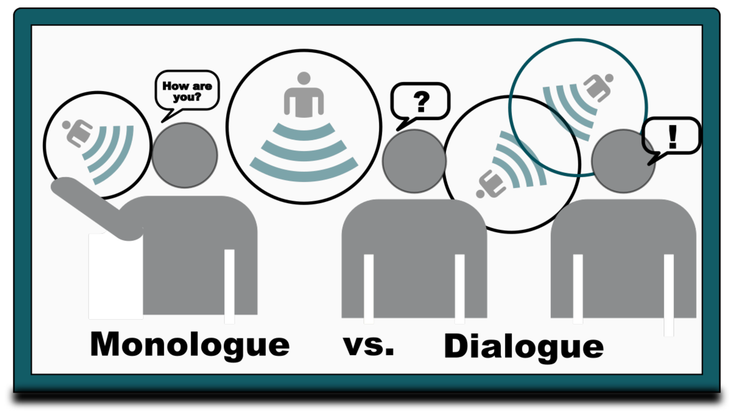

Fig. 1: Audio-visual stimulus with competing monologue vs. dialogue & multi-talker babble noise.

Research Questions (RQ):

1. **Is it possible to decode attended dialogues from single-trial EEG responses?**
2. **Are the neural responses to dialogues different from those to monologues?**

Methods

Participants:

- N: 17 (7 ♀)
- M_{age}: 26 (± 3.54)
- Normal hearing
  PTA4 ≤ 20 dB HL
- Native Danish
- No/corrected visual impairment

Stimulus:

- Audio-visual monologue vs. dialogue recordings
  Level = 60 dB SPL, TMR = 0 dB
- Multi-talker babble noise
  SNR = 5 dB
- Difficulty rating after each trial
  no significant subjective acoustic difference between mono and dia condition
- 3-AFC question after each trial
  average %-correct (main): 80 % for both mono and dia condition

Setup:

- **EEG:** BioSemi ActiveTwo
  64 cap + 2 mastoid electrodes; fs_{EEG} = 8192 Hz
- **Audio:** Loudspeakers equalized at listening position
  fs_{Audio} = 48 kHz
- **Eye-Tracking:** Tobii Pro Glasses II & Vicon
  fs_{Tobii} = 50 Hz, fs_{Vicon} = 100 Hz
- **Room:** Acoustically treated lab
  $T_{20} = 0.3$ s

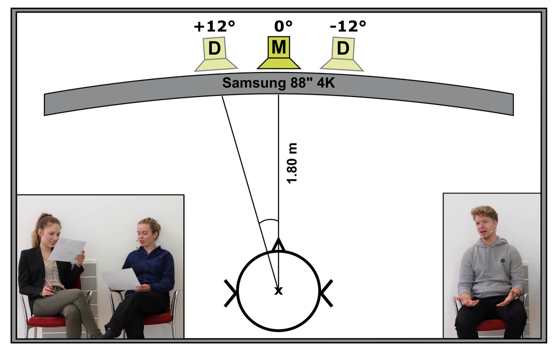

Fig. 2: Experimental setup for baseline measurement using either dialogue (left / D) or monologue (right / M) condition only.

Baseline:

10 trials of 120 s
5 x mono, 5 x dia (**only**)
No background noise

Main:

24 trials of 120 s
12 x mono, 12 x dia (**attended**)
Multi-talker babble noise

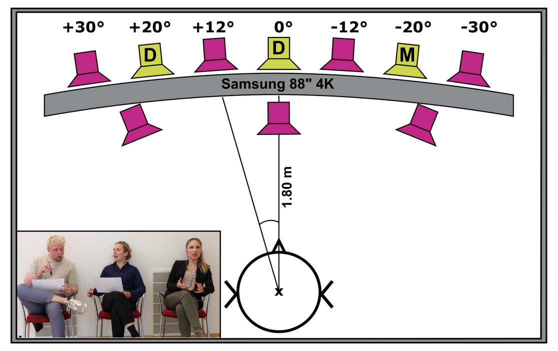

Fig. 3: Experimental setup for main measurement using dialogue vs. monologue condition with additional multi-talker babble noise.

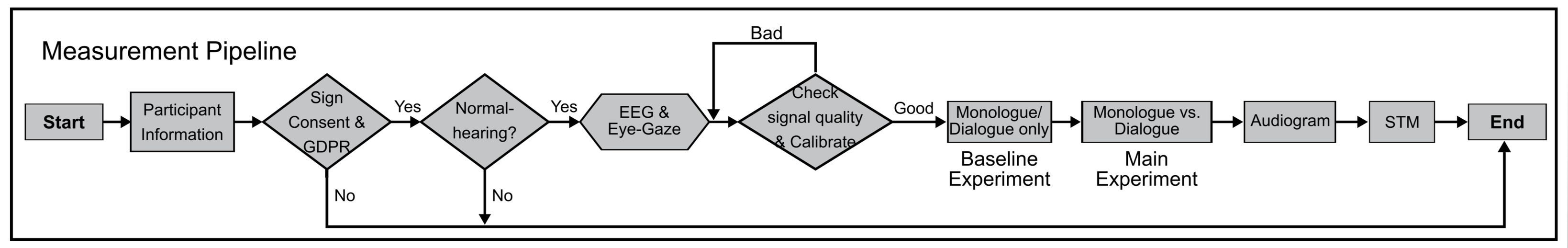

Fig. 4: Flow-chart of the measurement pipeline describing the entire experimental procedure.

Analysis

Gaze-Analysis:

- Quality scoring
- Drift / noise removal
- Velocity threshold
  (VT = 70 deg/s)
- Averaged saccade amplitude & number

Forward-Model:

- mTRF toolbox
  (Crosse et al. 2016)
- Estimate temporal response functions (TRFs)
- Wide range of positive time lags
  ($-100\ \mathrm{ms} < \tau < 500\ \mathrm{ms}$)
- 1 Layer cross-validation
  ($\lambda = 10^{0\ldots12}$)

Backward-Model (Classification):

- Estimate audio envelope
- Pearson's $\mathbf{r}$ based classification
- Wide range of negative time lags
  ($-500\ \mathrm{ms} < \tau < 100\ \mathrm{ms}$)
- Generic training
  (all – dia/mono)
- Specific training
  (dia/mono – dia/mono)

Tikhonov regularization: $\hat{\beta}_R = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$

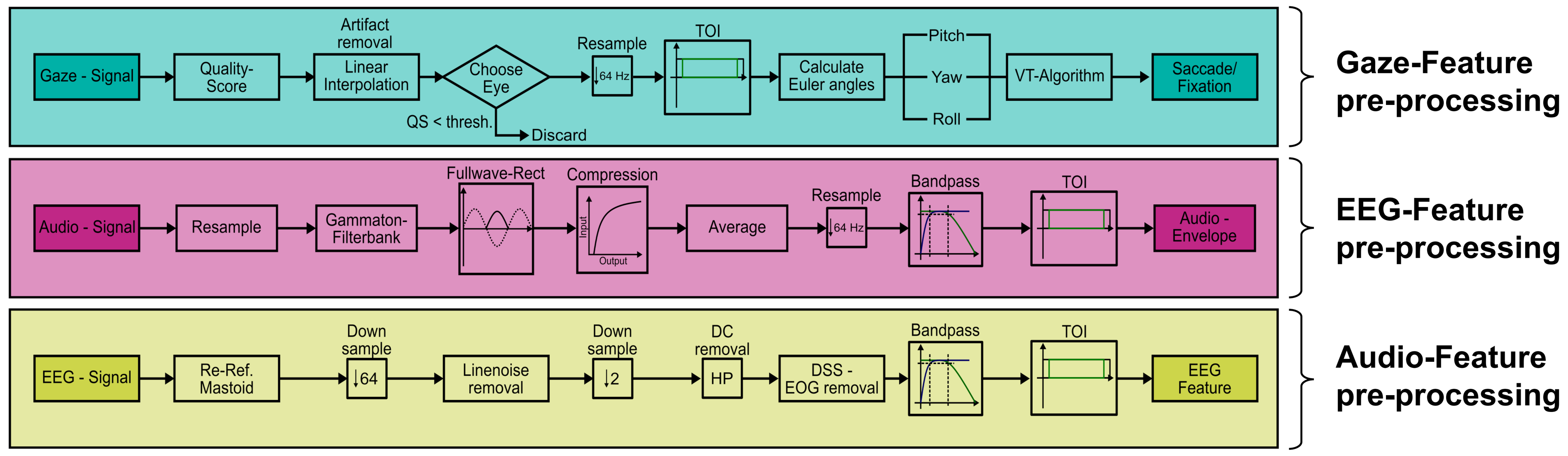

Fig. 5: Flow-chart of the analysis pipeline (inspired by Fuglsang et al. 2017) to pre-process relevant data streams e.g. Audio-Envelope (red), EEG-Response (yellow) and Gaze-Signal (green.)

Results

Gaze analysis:

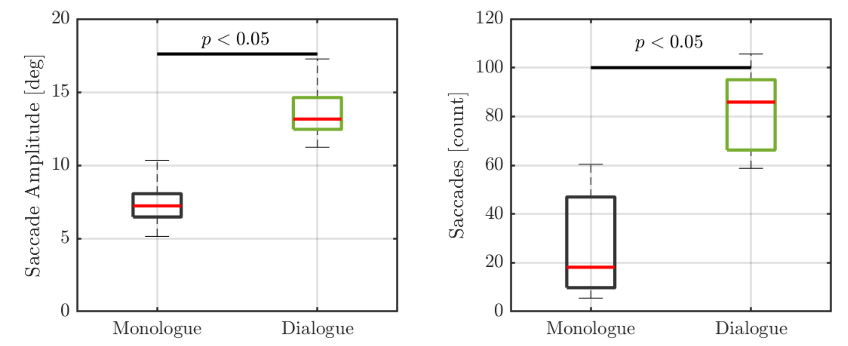

Fig. 6: Subject averaged saccade amplitude for monologue and dialogue atttended condtion (left). Number of performed saccades averaged across subjects for each condition (right). Saccade identification is based on the configuration of the VT-Algorithm.

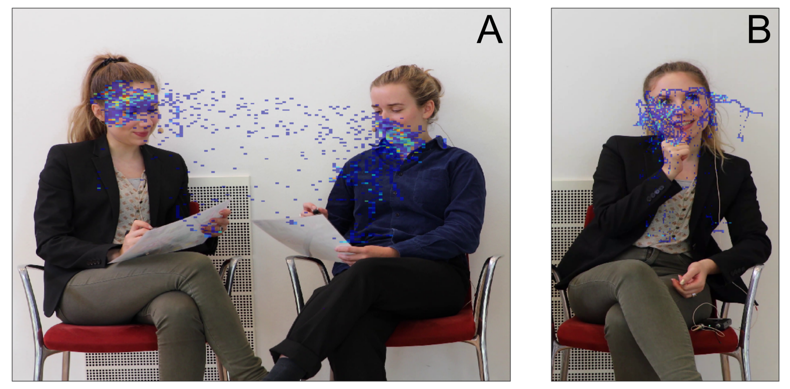

Fig. 7: Example heatmap of the fixation points in dialogue (A) and monologue (B) attended condition. The red-ish colour indicates multiple fixations on one spot.

Forward Model results:

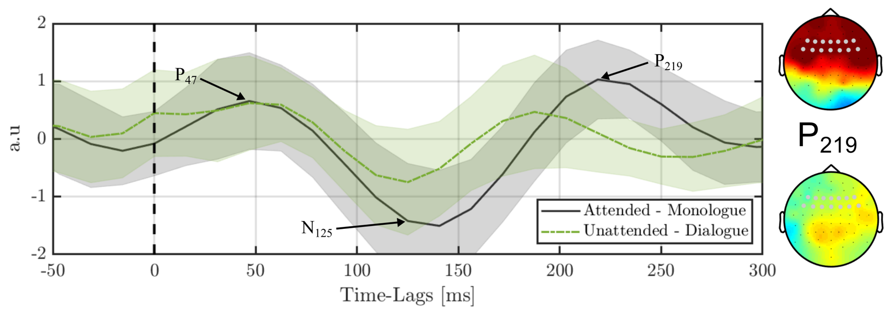

Fig. 8: Temporal response function at time-lags between -50 ms and 300 ms post-stimulus. The subjects' task was to attend to the monologue and ignore the dialogue. The arrows P_{47}, N_{125} and P_{219} mark prominent peaks and dips. The topology plot on the right shows the weight distribution across electrodes for the attended (top) and unattended (bottom) case at 219 ms post-stimulus (P219).

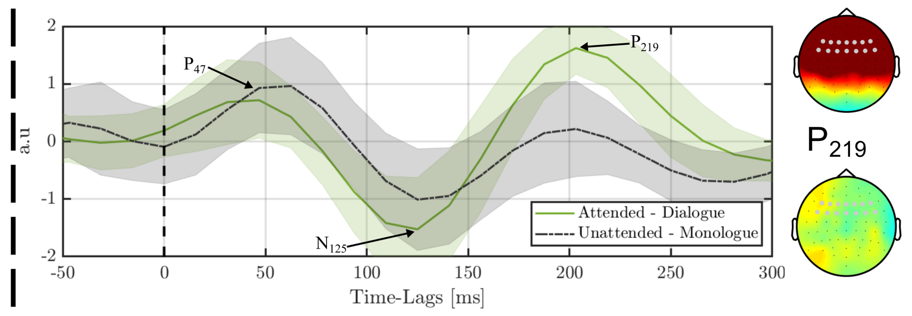

Fig. 9: Temporal response function at time-lags between -50 ms and 300 ms post-stimulus. The subjects' task was to attend to the dialogue and ignore the monologue. The arrows P_{47}, N_{125} and P_{219} mark prominent peaks and dips. The topology plot on the right shows the weight distribution across electrodes for the attended (top) and unattended (bottom) case at 219 ms post-stimulus (P219).

Classification results:

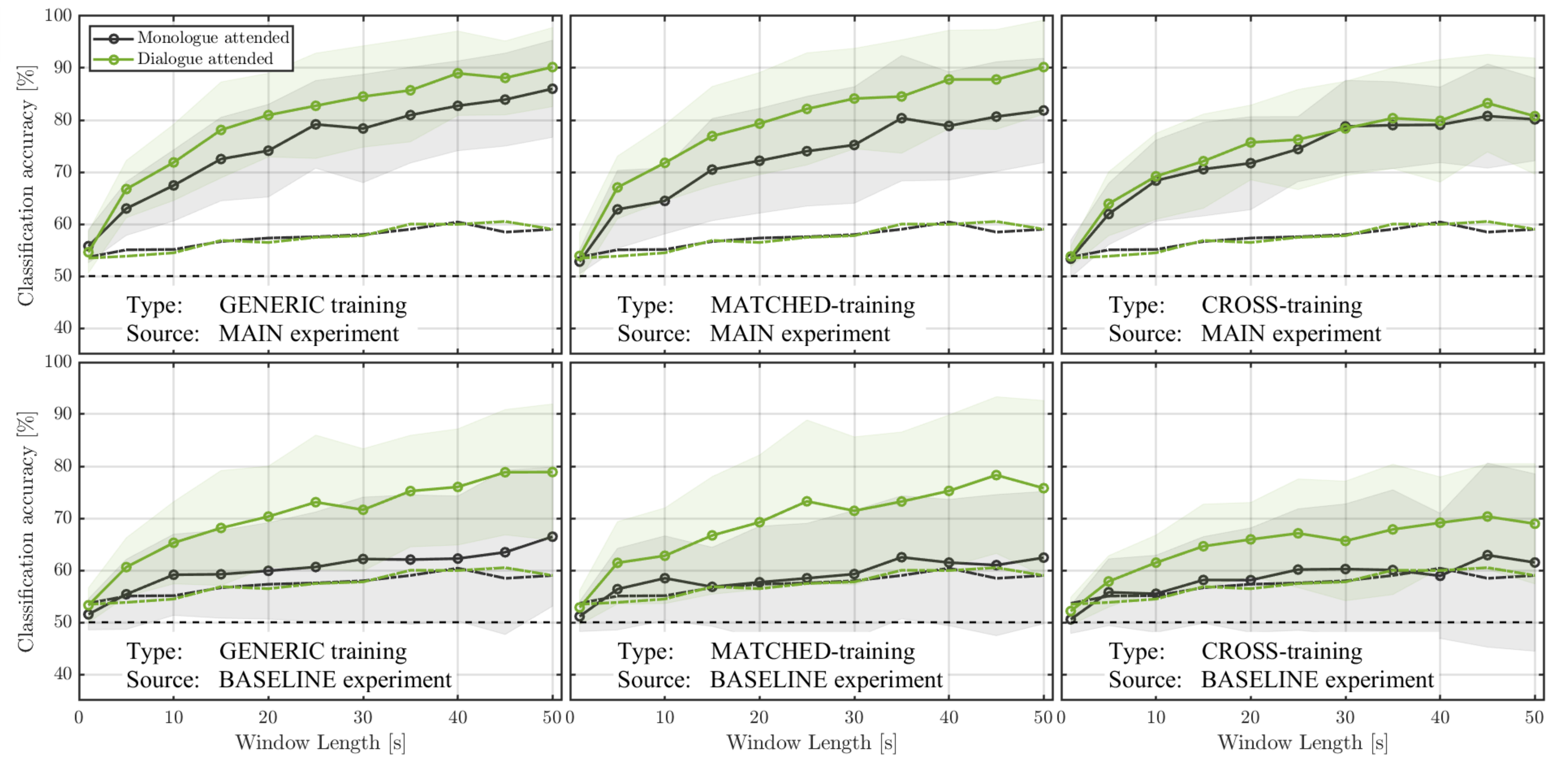

Fig. 10: Average classification accuracies obtained for all monologue and dialogue attended trials during the main experiment. The shaded areas represent +/- 1 std, the dash-dotted lines the noise floors (95th percentile of classification accuracies with phase-scrambled data). Data from the main experiment (top row) and the baseline experiment (bottom row) were used to train attended backward models. Left column: generic training with mono and dia attended trials; middle column: object-specific matched-training (mono/mono, dia/dia); right column: object-specific cross-training (mono/dia, dia/mono).

Discussion

- The monologue vs. dialogue audio-visual stimulation paradigm induces:
  - No significant differences in behavioural scores (comprehension and difficulty ratings)
  - Clear differences in eye-gaze behaviour (saccade amplitude and frequency)
  - Delayed and broadened TRFs in both conditions
  - Increased TRFs around 200 ms post-stimulus for the dialogue attended condition
- The EEG-based classification results show:
  - Highest classification accuracy when training data originates from the same acoustic condition (MAIN) as the test signals
  - Higher classification accuracy when the dialogue was attended, provided that the reconstruction model training was at least partly based on this condition
  - Largest differences between monologue and dialogue attended classification accuracies when the models were trained with the BASELINE condition data (single speech stream)
- **RQ1:** It is indeed possible to decode attended dialogues from single-trial EEG responses
- **RQ2:** The neural responses to dialogues appear to be different from those to monologues in that they are more pronounced
- The apparent differences between monologue and dialogue attended conditions could be related to
  - Higher listening effort expended when attending to a dialogue as indicated by Jaeger et al. 2020
  - Remaining eye-movement artifacts that are correlated with the acoustic speech features

Information

Sascha Bilert & Johannes Zaar
Eriksholm Research Centre
sascha_bilert@icloud.com
jozr@eriksholm.com

Eriksholm Research Centre
Rørtangvej 20
DK - 3070 Snekkersten
Phone +45 4829 8900

Read more at:
www.eriksholm.com:

In collaboration with:

DTU Health Technology
Department of Health Technology

EriksholmResearchCentre
PART OF OTICON

References: **1.** O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., and Lalor, E. C. (2015). "Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG", Cerebral Cortex 25. **2.** Das, N., Bertrand, A., and Francart, T. (2018). "EEG-based auditory attention detection: boundary conditions for background noise and speaker positions", Journal of Neural Engineering 15. **3.** Fuglsang, S. A., Dau, T., and Hjortkjær, J. (2017). "Noise-robust cortical tracking of attended speech in real-world acoustic scenes", NeuroImage 156, 435–444. **4.** Crosse, M. J., Di Liberto, G. M., Bednar, A., and Lalor, E. C. (2016). "The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli", Frontiers in Human Neuroscience 10, 1–14. **5.** Jaeger, M., Mirkovic, B., Bleichner, M. G., and Debener, S. (2020). "Decoding the Attended Speaker From EEG Using Adaptive Evaluation Intervals Captures Fluctuations in Attentional Listening", Frontiers in Neuroscience 14, 1–16.