INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Effects of spectral and temporal modulation degradation on intelligibility and cortical tracking of speech signals

Ignacio Calderón De Palma[1], *Laura S. Lopez*[2], *Alejandro Lopez Valdes*[3]

[1]Radboud University Medical Center, The Netherlands
[2]Washington State University, United States of America
[3]Trinity College Dublin; Global Brain Health Institute; Trinity College Institute of Neuroscience; Trinity Centre for Biomedical Engineering, Ireland
ignacio.calderondepalma@radboudumc.nl, laura.lopez@wsu.edu, alejandro.lopez@tcd.ie

Abstract

Understanding speech in challenging listening environments relies on diverse streams of information, including sensory signals, prior knowledge, and expectations. This is a challenge for patient populations who have compromised bottom-up sensory information. Neural entrainment evaluations can offer insights into the effects of signal degradation on speech processing. We collected electroencephalography from normal hearing listeners, in order to evaluate the effects of spectro-temporal information removal on speech intelligibility and cortical tracking. Our results showed a decrease in speech intelligibility with increased degradation and a decrease in encoding accuracy for the degraded conditions, compared to the clean control, but no differences in encoding accuracy between degradation conditions. We found significant differences between the weights of temporal response functions of clean and degraded speech conditions, which were specific to each type of degradation.

Index Terms: Neural tracking, speech intelligibility, spectro-temporal sensitivity, EEG

1. Introduction

Several studies have used cortical tracking in response to naturalistic stimuli providing insights into speech processing and remarking the impact of attention, prior information, and speech processing mechanisms [1-5]. Nevertheless, the specific consequences of reduced spectro-temporal modulation content on neural tracking can still be further addressed.

Spectro-temporal modulations are the fundamental building blocks of complex signals [6-7], carrying important cues for speech intelligibility [8]. Studying the limitations of missing spectral and temporal information are of particular importance for aging populations, as well as recipients of hearing devices as they receive less spectro-temporal information overall [9-10] or have a broader tuning [11], both detrimental for speech processing. Assessing cortical tracking under situations with spectral and temporal degradations can help us better understand how stimuli are encoded/decoded in these patient populations, while providing an objective measure to study the speech processing hierarchy, when confronted with bottom-up degradations in the input speech signal.

Currently, cortical tracking has contributed to our understanding of potential mechanisms associated with speech enhancement under complex listening scenarios. A previous study by Holdgraff and colleagues [12] showed that, in a repetition paradigm, context-specific information shifts the spectro-temporal tuning of neurons in the superior temporal gyrus, assessed using high-gamma activity recovered from electrocorticography recordings. In other words, context specific information provided an enhancement of the stimulus features. Using a similar paradigm Di Liberto et al. [13], also found influence of prior information, on low frequency cortical tracking, with influence mainly on phonetic feature encoding. Based on their results, the authors argued the possibility that the shifts in tuning to spectro-temporal information found in [12] could be the result of changes in phonemic encoding. Recently [14], proposed that envelope tracking is affected by both acoustic and speech specific processing, adding to the idea that envelope tracking is not only representative of bottom-up acoustic features, but that potentially, these top-down effects can permeate into lower processing stages. For degraded speech, encoding accuracy for electroencephalography (EEG) signals, using a spectro-temporal modulation model, has been shown to follow opposite effects with stimulus degradation, which depend on context [15]. Matching expectations leads to a decrease in accuracy as signals become clearer, while mismatched expectations lead to a decrease in encoding accuracy as signals become clearer.

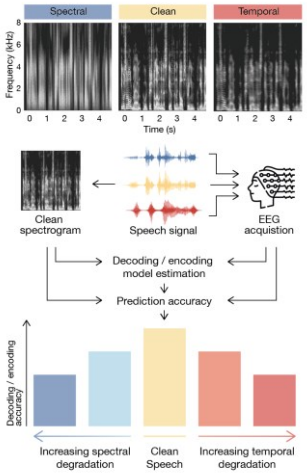

Figure 1: *Hypothesis and rationale for the study. Clean and degraded speech signals were presented to participants (top row shows degradations), while EEG signals were recorded. Encoding models were recovered using the clean speech signals. We show expectations for encoding prediction accuracy values.*

10.21437/Interspeech.2023-964

In this study, we collected a new EEG dataset in order to evaluate the effects of speech degradation on cortical tracking of acoustic features. We selectively removed spectro-temporal modulations from speech signals by filtering in the modulation domain, and keeping high levels of speech intelligibility (above ~60%). We expected that a decrease in modulation content would lead to a decrease in acoustic encoding accuracy (see Figure 1) showing the impact and dominance of acoustic features.

2. Methods

2.1. Participants

This study was approved by the research ethics committee of the School of Engineering at Trinity College Dublin and in compliance with the University's GDPR policies. Thirteen English speaking (native), right-handed participants (Age [range]: 25 [20-30] yrs., 8 female) volunteered for the study. Their hearing thresholds were assessed between 0.25 – 6 kHz and ensured to be below 20 dBHL (Hekto Diagnostic, Horentek Hearing Diagnostics, USA). All participants completed both behavioural and EEG assessments on the same day.

2.2. Stimulus degradation and delivery

Speech stimuli used a sampling frequency of 44.1 kHz, 16-bit resolution. The stimuli processing used in this study involved selective removal of spectral or temporal modulations using the procedure described in [8], also used in [12]. Briefly, signals were first transformed into their spectrogram representation using gaussian filters with 32 Hz width in the frequency domain (i.e., 5 ms in time domain). Next, the modulation power spectrum of the signal was obtained applying a two-dimensional Fourier transform. Low-pass filters in the modulation domain were then applied by zeroing coefficients beyond the filter cut-off. Finally, the stimuli were taken back to a time-frequency representation, via an inverse Fourier transform, which was inverted into a time waveform using an iterative procedure. All filtering was carried out offline.
Stimuli were delivered using insert earphones (Etymotic ER2, Etymotic Research Inc., USA) connected to a Tascam US-100 sound interface (TASCAM, US) and a FiiO Alpen headphone amplifier (Guangzhou FiiO Electronics Technology Co., Ltd., China).

2.3. Behavioural assessment

For the behavioural assessment, we used the first 15 lists of the IEEE sentence material [16]. Each of them consisted of 10 sentences, narrated by a male speaker. For each participant, we selected seven lists at random. Within each list, we randomized the modulation low-pass filters to be applied to each sentence. Based on pilot data, we selected spectral cut-offs of 0.15, 0.24, 0.39, 0.62 and 1 cyc/kHz and temporal cut-offs of 2, 3.3, 4.5, 5.8, 7 Hz, in order to obtain psychometric functions. All sentences were RMS normalized after filtering, and presented at a level of 63 dBA/channel.
The experiment was presented using MATLAB r2021b and was self-paced. Participants had to listen to each sentence and then type in what they perceived. Accuracy was assessed by evaluating the number of correctly identified unique words within each sentence.

2.4. EEG evaluation

2.4.1. Stimuli and task

For the material used in the EEG experiment, we selected 25 audio snippets from the audiobook of a classic novel ("The Old Man and the Sea"), narrated by a male speaker in English. Each of the snippets was approximately three minutes in length. The experiment had five conditions: two spectral and two temporal degradations, and a clean control. Each condition had an equal number of trials. We used the filtering procedure described in section 2.2, with temporal and spectral cut-offs of 5, 6 Hz and 0.3, 0.45 cyc/kHz, respectively. These limits were selected, based on a pilot study, in order to maintain intelligibility levels above 60%. All the snippets were RMS normalized after filtering, and presented at a level of 63 dBA/channel. Degradation and presentation order were randomized for each participant.
We presented our stimuli using Presentation software (Neurobehavioral Systems, Inc., USA), which provided a trigger sent via parallel port and recorded by the EEG amplifier. This allowed the correct synchronization between stimuli and EEG recordings.
EEG signals were recorded with a 64-channel BioSemi Active Two system (BioSemi B.V., The Netherlands) using a sampling rate of 1024 Hz. We recorded from 64 scalp electrodes and two reference channels located at the mastoids. Testing was carried out in a dark room and participants were instructed to maintain visual fixation on a cross centred on the screen for the duration of each trial. Participants were suggested to take breaks every three consecutive blocks and had a long break at trial number 15. The experiment was self-paced with further breaks allowed in between trials. To ensure attention, two questions were asked after each trial. Based on the presented snippet, participants could answer "True", "False" or "I don't know".

2.4.2. Feature extraction

We used spectrogram representations of the clean speech signals as the feature in our encoding analysis (Figure 1). To obtain this feature, we first filtered each of the clean speech audio files using a gammatone filterbank, consisting of 20 logarithmically spaced bands spanning the range 80-8000 Hz [17]. Next, we extracted the envelope by computing the signal power and then resampling the signal to 64 Hz, using a moving average filter (*mTRFenvelope* in [18]*).* Finally, we applied a compression to the envelopes by raising them to the power of $\log_{10}(2)$.

2.4.3. Pre-processing of EEG signals

EEG processing was carried out using Fieldtrip toolbox [19], NoiseTools [20] and custom-made code, in Matlab r2021b.To summarize, the EEG signals were initially re-referenced to the mastoids. This is an obligatory step for BioSemi devices, in order to achieve a correct common mode rejection. Next, we low-pass filtered (Butterworth, 3rd order, zero-phase shift) the signals with a cut-off of 8 Hz and down-sampled to 64 Hz. We then applied a robust re-reference method (*nt_detrend* in Noise Tools), to get to an average reference and to remove step-like responses if present. After this step, we applied a high pass filter at 1 Hz (Butterworth, 3rd order, zero-phase shift), which completed the bandpass filter procedure, targeting delta and theta bands (1-8 Hz). We subsequently identified poor quality channels as those whose median standard deviation was below 1/6 or above 3 times the median standard deviation across all

channels. These channels were interpolated using a spherical spline interpolation based on the time series of its neighbours. Finally, we performed independent component analysis, with the number of independent components being the rank of the EEG data matrix. This step was only used to project-out eye-blink components, which were determined by visual inspection of the time-series and topographic plots. The data was then epoched to the trial length and again re-referenced to the average between the mastoids.

2.4.4. Multivariate temporal response function (mTRF) fitting

To get the mTRFs for each condition, we used the mTRF Toolbox [18]. We used regularized (ridge) regression between the spectrogram features and the EEG data, with a time window from -150 to 450 ms. We applied both regularization ($\lambda=10^{-8}$, 10^{-7},…, 10^{8}) and used a leave-one-out cross-validation between trials, to avoid over-fitting and maximize prediction accuracy. Accuracy was always assessed using Pearson correlation (r). Once the data was fitted, we obtained for each participant and condition a null distribution of prediction accuracies, achieved by obtaining prediction accuracies with 100 random permutations of trials as well as circular-shifting the features.

We assessed prediction accuracy for every condition, by fitting models for each participant, using clean speech as a predictor and the weights extracted from the clean speech condition for 15 fronto-centrally distributed electrodes. For each participant, we calculated the median mTRF weights for the clean condition trials, after z-scoring the weights for each trial. Then we used this clean set of model weights and clean speech inputs, to predict EEG responses in all conditions. We report the average encoding accuracy across the selected electrodes.

We performed a separate encoding analysis to obtain model weights for every condition, including their prediction accuracy and null distributions, for each participant, electrode, and trial. We intended to evaluate differences between model weights when the predictor was the clean speech signal. We obtained mTRFs for each condition, by averaging only those electrodes above significance level within trials with mean accuracy above significance level. Given that we allowed the model weights to be adjusted across conditions, encoding accuracy was typically above significance. However, we had to exclude subjects 11 and 13 from two conditions and subject 2 from one condition. Finally, we obtained normalized TRFs for each participant by calculating their z-scores across trials for each condition and calculated the grand-average weights across participants as dimensionless mean z-scores (additional material).

3. Results

3.1. Behavioural assessment

Removing spectral or temporal information caused a drop in intelligibility (Figure 2). We noted that performance typically reached a plateau, forcing us to adjust a lapse rate of 10%. These results are in-line with other studies using similar filtering techniques [8, 21].

At the group level, the cut-offs selected for temporal and spectral degradation for the EEG evaluation roughly correspond to the same intelligibility levels for both degradations (*correct word discrimination: mean ± σ: Temporal<5 Hz: 74% ± 3.4%, Spectral<0.3 cyc/kHz: 68% ± 4.5%, Temporal<6 Hz: 87% ± 1.2%, Spectral<0.45 cyc/kHz: 83% ± 2.7%*). These are still high levels of speech intelligibility, though we found them to be significantly different from each other (*paired samples t-test between the means at each level of degradation, extracted from logistic fit Spectral: t(12)=-6.54, p<0.01, Temporal: t(12)=-6.02, p<0.01*).

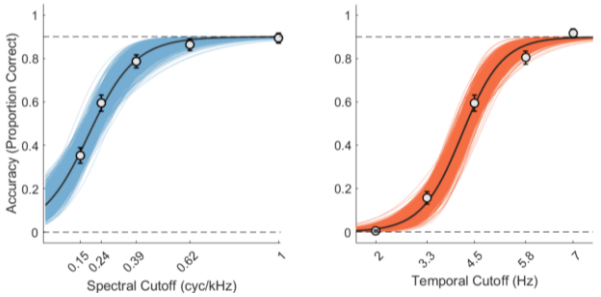

Figure 2: *Group level intelligibility outcomes for spectral (left) and temporal (right) degradation. Dots represent the mean and error bars the 95% confidence intervals for pooled data. Logistic fits are included (dark trace represents the mean).*

3.2. Encoding models

We tested our main hypothesis by evaluating the decoding and encoding accuracy from our degraded stimulus paradigm. Firstly, in line with our hypothesis, prediction accuracy dropped when using the clean speech model, meaning that stimulus degradation was evident in the EEG signals (Figure 3).

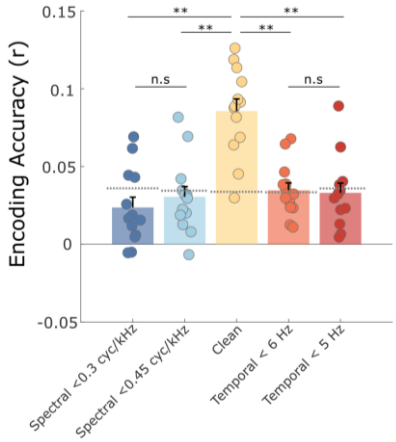

Figure 3: *Group level encoding accuracy. Error bars correspond to standard error around the mean accuracy. Each dot represents the average for a participant. Gray horizontal lines correspond to significance levels. Asterisks mark significant differences (p<0.01).*

This was further assessed with a repeated measures ANOVA, showing main effects for condition (*$F(4,48)=38.88$, $p<0.01$, $\eta^2=0.764$*). However, this difference did not follow the gradual trend we had foreseen, dropping to a similar level across conditions, which was typically below significance level (*paired samples t-test between conditions and significance level, Clean: t(12)=6.47, p<0.01; Temporal<5 Hz: t(12)=-0.28, p=0.78; Temporal<6 Hz: t(12)=0.11, p=0.91; Spectral<0.3 cyc/kHz: t(12)=-1.66, p=0.12; Spectral<0.45 cyc/kHz: t(12)=-0.76, p=0.46*). Post-hoc comparisons, Bonferroni corrected, showed that the clean speech condition was the only condition significantly different from the others (*Clean vs. Temporal<5 Hz: t(12)=-9.26, p<0.01; Clean vs. Temporal<6 Hz: t(12)=8.97, p<0.01; Clean vs. Spectral<0.3 cyc/kHz: t(12)=10.92, p<0.01;*

Clean vs. Spectral<0.45 cyc/kHz: t(12)=9.71, p<0.01) and there were no significant differences between the other conditions.
The evaluation of the weights of the encoding models, obtained for each condition but using clean speech as a predictor, showed that the clean speech condition (Figure 4, top panel) was in line with previous literature using comparable stimuli and paradigm [4]. In an exploratory analysis, we evaluated whether differences between neural responses for each condition were identifiable in the weights of the mTRFs. We performed a cluster-based non-parametric permutation analysis (N=5000, corrected for multiple comparisons), to evaluate differences ($p<0.05$) between the weights obtained for the degraded conditions and the clean control (Figure 4, bottom panel). We found significant differences at similar latency ranges for both types of degradations, around 100-200 ms, but for different frequency regions in each condition.

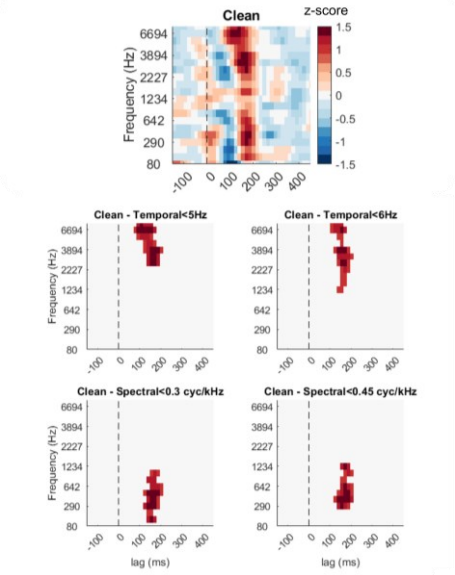

Figure 4: *Top: Weights of encoding model for the clean speech condition. Bottom: Results of permutation-based analysis. Shaded regions correspond to significant differences (p<0.05) where Clean>Degraded.*

4. Discussion

Behavioural results were in-line with the notion that spectral and temporal modulations are key features for speech understanding. The behavioural outcomes for both conditions also followed a previous study that used the same filtering procedure [8]. For the temporal degradation condition, our participants performed slightly better than those from another study [21], however, we note that the material and filtering techniques used differed from ours. For the spectral degradations, another study [12] used the same technique we used but, employed a cut-off at 0.5 cyc/kHz to produce unintelligible speech. Here we needed stricter cut-offs to achieve a comparable reduction in speech intelligibility, with our results being in-line with the notch-filtering results in [8].
The results of the encoding analysis followed our hypothesis, showing that prediction accuracy was affected by the stimulus degradation. This was evidenced as a decrease in prediction accuracy, which has been observed in previous studies [14, 15]. However, we did notice that prediction accuracy dropped for most subjects, being on average, below significance level and did not follow the gradual trend we considered. It is likely that the degradation was high enough to get to this low accuracy levels, but also similar intelligibility across conditions should be considered. In [14], the authors pointed out that in general, envelope encoding is dominated by acoustics, however, intelligibility also affects the encoding/decoding accuracy. In our case, it is possible that the intelligibility difference between conditions was small enough to be over-run by the applied degradations, without significant contribution to the overall encoding accuracy. An analysis of the sensitivity of mTRFs to spectro-temporal degradations with intelligibility changes could be considered using further degradation levels.
When comparing the degraded stimulus envelopes against the clean speech envelope, we observed a decrease in correlation coefficients between the clean and degraded envelopes (additional material). We saw no shifts in tuning, as mentioned in [12], interpreted here as an increase in weights for a specific frequency band, but rather a drop in the weights around the frequency bands where the degradation was the strongest. We interpret these differences between weights to be indexing the bottom-up degradation we applied; thus, it is possible that within this framework, responses were driven primordially by acoustics. The differences observed between the clean and degraded mTRFs, show consistent variations across conditions. The latency where differences were observed was around 100-200 ms for both types of degradation. For the spectrally degraded condition we saw a reduction for the lower-mid frequencies (80-1300 Hz). Part of this lower frequency range is typically where first and second formants are found [22] and is a region carrying pitch information [8]. On the other hand, for the temporal modulations, we observed effects at similar latencies, but for higher frequencies (above 1200 Hz). These are frequency bands associated with consonant information. Whether a reduction in mTRF weights, with reduced bottom-up information, can be an objective indicator of the impact of speech degradation on components of speech remains to be addressed, for example, by evaluation of phonemic loss in the behavioural data and extending our analysis using linguistic models.

5. Conclusion

In this study, we collected an EEG dataset using continuous stimuli consisting of clean and spectro-temporally degraded speech via modulation filters whilst maintaining high intelligibility. This allowed us to evaluate multivariate temporal response functions in normal hearing individuals when confronted with reduced bottom-up information. Our framework is an effort to understand how selective loss of spectral or temporal modulation information would manifest in linear encoding models. We found that our stimulus manipulations were evidenced in cortical tracking. With each type of degradation having characteristic and significantly different weights from the clean speech condition at specific frequency ranges, within the same time window. Future research will evaluate the mTRFs at a higher step in the speech processing hierarchy, addressing differences between acoustic and linguistic models, and including a thorough analysis of phonemic loss across participants.

6. Acknowledgements

This project has received funding from the European Union's Horizon 2020 framework program for research and innovation under the Marie Sklodowska-Curie grant agreement No 860718 (MOSAICS).

7. References

[1] N. Ding, J.Z. Simon. “Emergence of neural encoding of auditory objects while listening to competing speakers”. Proceedings of the National Academy of Sciences, vol. 109, no. 29, pp. 11854-11859, Jul. 2012.

[2] J.E. Peelle, M.H. Davis. “Neural Oscillations Carry Speech Rhythm through to Comprehension”. Frontiers in Psychology. vol 3, no. 320, pp 1-17, Sep. 2012.

[3] J.A. O'Sullivan, A.J. Power, N. Mesgarani, S. Rajaram, J.J. Foxe, B.G. Shinn-Cunningham, M. Slaney, S.A. Shamma, E.C. Lalor. “Attentional Selection in a Cocktail Party Environment can be decoded from single-trial EEG”. Cerebral Cortex, vol. 25, no 7, pp. 1697-1706. Jul. 2015.

[4] G.M. Di Liberto, J.A. O'Sullivan, E.C. Lalor. “Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing”. Current Biology, vol. 25, no. 19, pp. 2457-2465. Sep. 2015.

[5] E.S. Teoh, F. Ahmed, E.C. Lalor. “Attention Differentially Affects Acoustic and Phonetic Feature Encoding in a Multispeaker Environment”. J. Neuroscience, vol. 42, no. 4, pp. 682-691. Jan. 2022.

[6] T. Chi, Y. Gao, M.C. Guyton, P. Ru, S. Shamma. “Spectro-temporal modulation transfer functions and speech intelligibility”. J Acoustical Society of America, vol. 106, no. 5, pp. 2719-2732, Nov. 1999.

[7] N.C. Singh, F.E. Theunissen. “Modulation spectra of natural sounds and ethological theories of auditory processing”. J Acoustical Society of America, vol. 114 no. 6, pp. 3394-3411. Dec. 2003.

[8] T.M. Elliott, F.E. Theunissen. “The modulation transfer function for speech intelligibility”. PLoS Computational Biology, vol. 5, no. 3, e1000302, Mar 2009.

[9] M.B. Winn, J.H. Won, I.J. Moon. “Assessment of Spectral and Temporal Resolution in Cochlear Implant Users Using Psychoacoustic Discrimination and Speech Cue Categorization”. Ear & Hearing, vol. 37, no. 6, pp. 377-390, Nov. 2016.

[10] L.C.E. Veugen, A.J. van Opstal, M.M. van Wanrooij. “Reaction Time Sensitivity to Spectrotemporal Modulations of Sound”. Trends in Hearing, vol. 26, Sep. 22.

[11] J. Erb, L.M. Schmitt, J. Obleser. “Temporal selectivity declines in the aging human auditory cortex”. Elife, vol. 3, no 9, e55300, Jul. 2020.

[12] C.R. Holdgraf, W. de Heer, B. Pasley, J. Rieger, N. Crone, J.J. Lin, R.T. Knight, F.E. Theunissen. “Rapid tuning shifts in human auditory cortex enhance speech intelligibility”. Nature Communications, vol 7 (13654), Dec. 2016.

[13] G.M. Di Liberto, M.J. Crosse, E.C. Lalor. “Cortical Measures of Phoneme-Level Speech Encoding Correlate with the Perceived Clarity of Natural Speech”. eNeuro, vol. 5, no. 2, Apr. 2018.

[14] K.D. Prinsloo, E.C. Lalor. “General Auditory and Speech-Specific Contributions to Cortical Envelope Tracking Revealed Using Auditory Chimeras”. J Neuroscience, vol 42, no. 41, pp. 7782-7798. Oct. 2022.

[15] E. Sohoglu, M. H. Davis. “Rapid computations of spectrotemporal prediction error support perception of degraded speech”. Elife, vol. 4, no. 9, e58077, Nov. 2020.

[16] “IEEE Recommended Practice for Speech Quality Measurements”, in IEEE Transactions on Audio and Electroacoustics, vol. 17, no. 3, pp. 225-246, Sep 1969.

[17] F. Apoux, E.W. Healy. “On the number of auditory filter outputs needed to understand speech: further evidence for auditory channel independence”. Hearing Research, vol. 255, no. 1-2, pp. 99-108. Sep. 2009.

[18] M.J. Crosse, G.M. Di Liberto, A. Bednar, E.C. Lalor. “The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli”. Frontiers in Human Neuroscience, vol. 10 , Nov. 2016.

[19] R. Oostenveld, P. Fries, E. Maris, J.M. Schoffelen. “FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data”. Computational Intelligence and Neuroscience, vol. 2011, id. 156869, Dec. 2010.

[20] A. de Cheveigné, D. Arzounian. “Robust detrending, rereferencing, outlier detection, and inpainting for multichannel data”. Neuroimage, vol. 15, no. 172, pp. 903-912, May 2018.

[21] A. Flinker, W.K. Doyle, A.D. Mehta, O. Devinsky, D. Poeppel. “Spectrotemporal modulation provides a unifying framework for auditory cortical asymmetries”. Nature Human Behaviour, vol. 3, no. 4, pp. 393-405. Apr. 2019.

[22] R.L. Diehl. “Acoustic and auditory phonetics: the adaptive design of speech sound systems”. Philos. Trans. of the Royal Society London B Biological Sciences, vol 363, pp. 965-978. Mar. 2008.