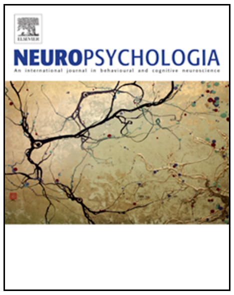
Author(s): Hämäläinen, Jarmo; Parviainen, Tiina; Hsu, Yi-Fang; Salmelin, Riitta

Title: Dynamics of brain activation during learning of syllable-symbol paired associations

Hämäläinen, J., Parviainen, T., Hsu, Y.-F., & Salmelin, R. (2019). Dynamics of brain activation during learning of syllable-symbol paired associations. Neuropsychologia, 129, 93-103. https://doi.org/10.1016/j.neuropsychologia.2019.03.016

Accepted Manuscript


NEUROPSYCHOLOGIA
An international journal in behavioural and cognitive neuroscience

Dynamics of brain activation during learning of syllable-symbol paired associations

Jarmo A. Hämäläinen, Tiina Parviainen, Yi-Fang Hsu, Riitta Salmelin

PII: S0028-3932(18)30613-4
DOI: https://doi.org/10.1016/j.neuropsychologia.2019.03.016
Reference: NSY 7053

To appear in: *Neuropsychologia*

Received Date: 17 September 2018
Revised Date: 20 February 2019
Accepted Date: 25 March 2019

Dynamics of brain activation during learning of syllable-symbol paired associations

Jarmo A. Hämäläinen[a*], Tiina Parviainen[a], Yi-Fang Hsu[b,c], Riitta Salmelin[d,e]

[a] Centre for Interdisciplinary Brain Research, Department of Psychology, P.O. Box 35, 40014 University of Jyväskylä, Finland

[b] Department of Educational Psychology and Counseling, National Taiwan Normal University, 10610 Taipei, Taiwan

[c] Institute for Research Excellence in Learning Sciences, National Taiwan Normal University, 10610 Taipei, Taiwan

[d] Department of Neuroscience and Biomedical Engineering, 00076 Aalto University, Finland

[e] Aalto NeuroImaging, 00076 Aalto University, Finland

*Corresponding author:

Jarmo Hämäläinen

Department of Psychology

P.O. Box 35

40014 University of Jyväskylä

Finland

Phone: +358 40 8053490

Email: jarmo.a.hamalainen@jyu.fi

Abstract

Initial stages of reading acquisition require the learning of letter and speech sound combinations. While the long-term effects of audio-visual learning are rather well studied, relatively little is known about the short-term learning effects at the brain level. Here we examined the cortical dynamics of short-term learning using magnetoencephalography (MEG) and electroencephalography (EEG) in two experiments that respectively addressed active and passive learning of the association between shown symbols and heard syllables. In experiment 1, learning was based on feedback provided after each trial. The learning of the audio-visual associations was contrasted with items for which the feedback was meaningless. In experiment 2, learning was based on statistical learning through passive exposure to audio-visual stimuli that were consistently presented with each other and contrasted with audio-visual stimuli that were randomly paired with each other. After 5 to 10 minutes of training and exposure, learning-related changes emerged in neural activation around 200 and 350 ms in the two experiments. The MEG results showed activity changes at 350 ms in caudal middle frontal cortex and posterior superior temporal sulcus, and at 500 ms in temporo-occipital cortex. Changes in brain activity coincided with a decrease in reaction times and an increase in accuracy scores. Changes in EEG activity were observed starting at the auditory P2 response followed by later changes after 300 ms. The results show that the short-term learning effects emerge rapidly (manifesting in later stages of audio-visual integration processes) and that these effects are modulated by selective attention processes.

Highlights

- MEG and EEG were recorded during audio-visual training and exposure
- Changes in brain activity emerged 5 - 10 min after learning
- During passive exposure changes emerged first at 200 ms
- Late phases of audio-visual integration were also affected (350 ms)
- Active training utilizes frontal cortex during training

Keywords: audio-visual, electroencephalography, learning, magnetoencephalography

Acknowledgements

This work was supported by the Academy of Finland (profiling action ‘MultiLeTe’ grant number #292466; personal grants #255349, #256459, #283071 and #315553 to RS), the European Union H2020 MSCA-ITN-2014-ETN Programme, “Advancing brain research in children’s developmental neurocognitive disorders” -project (ChildBrain, grant number 641652), Taiwan Ministry of Science and Technology (grant number MOST105-2410-H-003-145-MY3 and MOST107-2636-H-003-001 to YFH), and Sigrid Jusélius Foundation funding (to RS). The authors declare no competing financial or other interests.

1. Introduction

Despite the ease with which we learn to associate events across the senses, relatively little is known about the immediate learning processes in the human brain that occur at the beginning stages of training of cross-modal associations. Most of the studies examining the acquisition of such associations have examined long-term learning occurring in the time span of months (e.g., Maurer et al., 2008, Brem et al., 2010).

While studies examining the long-term learning effects have been important in establishing the brain mechanisms involved in cross-modal processing, it is not known which of these brain mechanisms are used during the initial steps of the learning process, and if there are distinct stages of learning during which some of the mechanisms are more important than others. Current theoretical models suggest a reciprocal information transfer between areas processing the visual form and auditory features manifested in fast sensory brain responses, as well as the sensory association areas at later stages in brain responses reflecting extended phases of cross-modal integration (Calvert, 2001; Bernstein & Liebenthal, 2014; Murray et al., 2015). Attention and working memory systems are also thought to assist in the learning process (Calvert, 2001; van Atteveldt et al., 2009).

Studies on long-term effects of audio-visual learning provide a starting point for expected short-term learning effects. Audio-visual processing engages specific cross-modal sites and primary sensory areas (e.g., Raij et al., 2000; van Atteveldt et al., 2004; Molholm et al., 2006; Kayser et al., 2009; Murray et al., 2015). The superior temporal sulcus in the left hemisphere has been implicated particularly in processing of well-established letter-speech sound combinations, thus mostly reflecting long-term audio-visual memory representations (Raij et al., 2000; van Atteveldt et al., 2004; Hashimoto & Sakai, 2004, Blomert, 2011). The fusiform gyrus in the left hemisphere is particularly involved in learning letters (Cohen et al., 2000) but is sensitive to long-term learning occurring over several months (e.g., Maurer et al., 2008; Brem et al., 2010). Long-term exposure to grapheme-phoneme associations also affects perception of ambiguous speech through mechanisms in the posterior superior temporal cortex and inferior parietal lobe (Bonte et al., 2017). This suggests that several interacting processes are affected by learning audio-visual associations and the processes most prominently activated depend on the task parameters and demands.

Training effects occurring within days to learning of audio-visual combinations have been found in the left parieto-temporal cortex and posterior inferior temporal gyrus (Hashimoto & Sakai, 2004; Karipidis et al., 2016). Additionally, the left occipito-temporal area, together with primary auditory and visual areas, showed increased activity during audio-visual learning (Tanabe et al., 2005). The important role of the posterior brain regions was also shown in a transcranial direct current stimulation study in which the learning rate and outcome of audio-visual associations was modulated by stimulation injected to left inferior parietal lobe (Younger & Booth, 2018). Furthermore, the frontal cortex has been suggested to mediate cross-modal learning, particularly in the case of arbitrary associations and sub-optimal presentation of cross-modal information, and to show activity change as a function of consistent audio-visual pairings (Gonzalo et al., 2000; Calvert, 2001; van Atteveldt et al., 2009).

Previous studies have also shown that overt attention can mask subtle effects of audio-visual processing that are observable only during implicit or passive presentation of the audio-visual material (van Atteveldt et al., 2004; Blau et al., 2009). In order to make the neural processes during learning of audio-visual associations observable at the cortical level, non-optimal presentation of the stimuli might therefore be needed (cf. van Atteveldt et al., 2004). In the current study, in an active learning task we employed non-synchronous presentation of the visual and auditory stimuli, while in a passive learning task we used synchronous presentation of the stimuli. Both active and passive tasks were used to examine possible general neural mechanisms related to learning of audio-visual associations.

Here the goal was to examine short-term changes in cortical dynamics in the brain regions thought to be engaged in audio-visual association learning. First, in an MEG experiment the effects of active audio-visual learning were examined on the processing of novel visual symbols. Second, in an EEG experiment the effects of passive exposure to simultaneously presented audio-visual pairings were examined. The two experiments should reveal if cross-modal learning effects can be observed during the training/exposure session. Further, they should reveal if there is a common time window during which changes in brain activity due to learning can be observed both with active feedback and using passive exposure. We expected to see changes in the neural activity of the sensory areas reflected in the low-level cortical responses at 100 - 200 ms after stimulus onset (Tanabe et al., 2005; Yoncheva et al., 2010). In addition, we expected to see modulation of activity at a later time window where cross-modal integration effects have been reported (Raij et al., 2000; Shams et al., 2005; Karipidis et al., 2016).

2. Material and methods

2.1 Experiment 1

2.1.1 Participants

Thirteen adult participants were included in the analyses (26.3 years on average, range 21-38 years; 7 female, 6 male; 12 right-handed, 1 ambidextrous based on self-report). From the total of 15 participants, one participant was excluded due to magnetic artifact from a tooth brace and one due to excessive eye blinks during the visual stimulus presentation. None of the participants had lived in Japan or studied Japanese (relevant for the choice of visual stimuli, see below). The experiment was carried out in accordance with the Declaration of Helsinki. The participants gave a written informed consent to participate. The study was approved by the Ethics Committee of the Aalto University.

2.1.2 Stimuli and experimental design

Auditory stimuli were recorded by a female native Finnish speaker in a sound-attenuated booth. The root-mean-square intensity level of the recorded consonant-vowel (CV) stimuli was adjusted to be equal. The CVs were /ka/, /ki/, /ko/, /pa/, /pi/, /pu/ (duration 176–266 ms). The auditory stimuli were presented at the level of approximately 75 dB through a panel speaker placed in front of the participant, above the head level. There were 6 visual stimuli modified from Japanese Hiragana writing symbols that were rotated 90 degrees to the right to avoid possible familiarity effects and false associations with their real corresponding syllables (Figure 1). The stimuli were dark grey on a light grey background, with the screen located approximately 1 m from the participant’s eyes. The symbols and CV syllables were each shown 120 times. The stimulus pairs were presented in a pseudorandom order such that the same symbol was never repeated immediately.

Each experimental trial started with a fixation cross shown at the centre of the screen for 500 ms (Figure 2). Thereafter, the visual symbol was presented for 1000 ms. The auditory syllable started playing at 600 ms after the onset of the visual stimulus. The participants were asked to judge whether the symbol was linked to the syllable or not. The delayed audio presentation was introduced in order to allow a clean access to cortical processing of the visual symbol without contamination by auditory activation, motor response, or response error monitoring. After the visual and auditory stimuli, a question mark appeared on the screen to prompt a response (max. 3000 ms): right index finger lift to indicate a match between symbol and syllable and right middle finger lift to indicate a mismatch. Immediately after the response, feedback appeared on the screen (the Finnish word for ‘correct’ or ‘incorrect’) for 1000 ms after which the next trial was presented. Accuracy and reaction time (with respect to question mark onset) were obtained for each trial. Due to technical difficulties, behavioral data from 3 participants was not available.

The focus of the study was on the modulation of the visual symbol processing as a result of learning the audio-visual associations. For this purpose two categories of trials were created, *learnable* and *non-learnable* (Figure 1). For half of the symbols, the feedback for the syllable associations was consistent throughout the experiment and the symbol-syllable link could thus be learned (*learnable category*). For the other half of the symbols the participants received the word ‘incorrect’ as the feedback and thus their association to syllables could not be learned (*non-learnable category*). This was to control for the general effect of exposure to the stimuli during the experiment and to separate the activity related to learning of the audio-visual associations.

Association learning was expected over the course of the experiment. In the beginning, participants’ responses were based on guessing; the correct combinations could be learned only through the feedback. Altogether, there were 360 trials in the *learnable* category and 360 trials in the *non-learnable* category leading to experiment length of approximately 40 minutes. In each category, each symbol was paired with each syllable 40 times.

2.1.3 Data recording and analysis

MEG data was collected using a 306-channel (102 magnetometers, 204 planar gradiometers) whole-head device (Elekta Oy, Finland) at the MEG Core of Aalto NeuroImaging, Aalto University, Finland. The data were high-pass filtered at 0.03 Hz, low-pass filtered at 200 Hz, and sampled at 600 Hz. The head position was monitored continuously using 5 small coils attached to the scalp (3 on the forehead and 2 behind the ears). Electro-oculography (EOG) was recorded using electrodes lateral to each eye (detection of horizontal eye movements) and above and below the left eye (blink detection).

Offline, head movements were corrected and external noise sources attenuated using the temporal extension of the signal space separation algorithm (Taulu et al., 2005) of the MaxFilter program (Elekta Oy, Finland). The head position was also converted to the default head position.

After the initial head movement correction the data was analysed using BrainStorm 3.2 (Tadel et al., 2011). Signal subspace projection was used to correct for eye blinks and horizontal eye movements. The further analyses used the planar gradiometers that are sensitive to brain activity directly under the sensor and are less sensitive to noise sources further away from the sensors compared to magnetometers (see Garcés et al., 2017 for the effect of channel choice in MEG). The MEG signal was low-pass filtered at 30 Hz and then segmented into trial-based time windows of -200 – 1000 ms with respect to the visual symbol onset (200 ms pre-stimulus baseline). Segments with over 3000 fT/cm peak-to-peak values were rejected.

The whole experiment was divided into quarters (1st set of 90 trials, 2nd set of 90 trials, 3rd set of 90 trials and 4th set of 90 trials for each category) and averaged by category (learnable, non-learnable). This allowed the examination of the initial stages of learning (1st quarter) and the stage where the audio-visual association was fully learned (4th quarter). After artifact rejection, 80 – 90 trials remained in each average.

A three-shell spherical head model fitted on the ICBM152 MRI template (Fonov et al., 2011) was used for calculating depth-weighted minimum norm estimates (wMNE) of the sources. The wMNE solution was restricted to the cortex. Noise covariance matrix was calculated from the baseline interval of the averaged responses.

The present analysis focused on the source activity, quantified as the absolute power, in a number of cortical areas. These were selected based on previous literature and examination of the grand-average source solutions as follows: The a priori regions of interest (ROIs) were based on the Desikan-Killiany (D-K) parcellation of the cortex (Desikan et al., 2006) and included sensory cortices in the left and right Heschl’s gyrus and lateral occipital cortex; cross-modal integration areas in the posterior superior temporal sulcus; attention and working memory related regions in the caudal middle frontal cortex and pars opercularis (Figure 3). Furthermore, the temporo-occipital area was included, outside of the D-K parcellation, because the grand average activity fell on the junction of multiple D-K parcels. The temporo-occipital ROI covered 200 vertices in each hemisphere.

Brain activity was examined in four consecutive 150 ms time windows encompassing activity from the basic visual evoked fields to the onset of the basic auditory evoked fields: 50 – 200 ms, 200 – 350 ms, 350 – 500 ms and 500 – 650 ms. Focus on the time window before the sound onset enabled exclusion of activity related to the motor decision, motor response and possible error monitoring processes.

2.1.4 Statistical analysis

Repeated measures ANOVAs (category [learnable, non-learnable] x quarter [1st, 2nd, 3rd, 4th] x hemisphere [left, right]) for each time window and region of interest were conducted. Effects involving interaction between category and quarter were of interest. Greenhouse-Geisser correction was applied when appropriate.

2.2 Experiment 2

2.2.1 Participants

Seventeen adult participants were included in the analyses (26.2 years on average, range 20-35 years; 14 female, 3 male; 16 right-handed, 1 left-handed based on self-report). From a total of 25 recruited participants, 8 participants were excluded; 5 due to excessive eye blinks during the visual stimulus presentation, 1 due to technical problem, and 2 due to low data quality. None of the participants had lived in Japan or studied Japanese (relevant for the choice of visual stimuli). The experiment was carried out in accordance with the Declaration of Helsinki. The participants gave a written informed consent to participate. The study was approved by the Ethics Committee of the University of Jyväskylä, Finland.

2.2.2 Stimuli and experimental design

Auditory stimuli were recorded by a female native Finnish speaker in a sound-attenuated booth. The root-mean-square intensity level of the recorded CV stimuli was adjusted to be equal. The CVs were /ko/, /to/, /ki/, /pi/, /pa/, /ta/ (duration 165 ms). The auditory stimuli were presented at the level of approximately 60 dB through a speaker placed approximately 1 m above the participant. As in Experiment 1, there were 6 visual stimuli modified from Japanese Hiragana writing symbols that were rotated 90 degrees to the right to avoid possible familiarity effects and false associations with their real corresponding syllables (Figure 4). The stimuli were dark grey on a light grey background, with the screen located approximately 1 m from the participant's eyes. The stimulus pairs were presented in a pseudorandom order such that the same symbol was never repeated immediately.

Each experimental trial started with a fixation cross shown at the centre of the screen for 745 ms. Thereafter, the visual symbol was presented for 700 ms. The auditory syllable started playing at the onset of the visual stimulus. The stimulus-onset asynchrony was 1450 ms. As a cover task the participants were asked to press a button when they saw a blue symbol or heard a syllable higher in pitch compared to the other syllables.

To examine the effect of association learning, two categories of trials were created, *learnable* and *non-learnable*. Half of the visual stimuli were always presented with its corresponding auditory stimuli (*learnable category*) while the other half of the visual stimuli were randomly paired with three auditory stimuli (*non-learnable category*). This comparison was to control for the general effect of exposure to the stimuli during the experiment and to separate the activity related to learning of the audio-visual associations.

Changes in brain activity were expected over the course of the experiment due to adaptation to the continuous stimulation and statistical learning effects. Altogether, there were 792 trials in the *learnable* category and 792 trials in the *non-learnable* category, leading to experiment length of approximately 40 minutes. Each stimulus was presented altogether 132 times.

2.2.3 Data recording and analysis

EEG data was collected using a 128-channel NeurOne amplifier (Bittium Oy, Finland) with Ag-AgCl electrodes attached to the HydroCel electrode net (Electrical Geodesics Inc., OR, USA) with Cz electrode as the reference. The data were high-pass filtered at 0.1 Hz, low-pass filtered at 200 Hz, and sampled at 1000 Hz. Electrode impedance was checked at the beginning of the recording and aimed to be below 50 kOhms for all channels. The data quality was additionally checked and monitored during the EEG recording.

The data was analysed using BESA Research 6.1 (BESA GmbH, Grafelfing, Germany). EEG was first examined for channels with poor data quality (mean: 4, range 0-10) that were rejected at this stage, and then segmented into trial-based time windows of -200 - 700 ms with respect to the visual symbol onset (200 ms pre-stimulus baseline). The whole experiment was divided into quarters (1st set of 198 trials, 2nd set of 198 trials, 3rd set of 198 trials, and 4th set of 198 trials for each category) and averaged by category (learnable, non-learnable). Segments with over 150 µV peak-to-peak values or 75 µV transient amplitudes were rejected (mean accepted number of trials: 142, range 58 - 181). After averaging, a low-pass filter of 30 Hz was applied, the channels with poor data quality were interpolated using the spherical spline method (Perrin et al., 1989), and the data was re-referenced to an average reference.

2.2.4 Statistical analysis

EEG data was then examined using cluster-based permutation tests (Maris & Oostenveld, 2007) in BESA Statistics 2.0. After initial t-test comparison between conditions of interest, the results were clustered based on time points and channels. Significance values for the clusters were based on permuted condition labels. The time window from 0 ms to 700 ms was used in the analysis. Cluster alpha of 0.05 was used with 3.5 cm channel neighbor distance and 3000 permutations. The learnable and non-learnable conditions were compared in each block.

3. Results

3.1 Experiment 1: Active learning

3.1.1 Behavioral results

All participants were able to learn the correct audio-visual associations during the first half (1st and 2nd quarters) of the MEG recording with only a few errors made after that. Accuracy was scored based on the response to the question “do the symbol and syllable form a pair” (for non-learnable items the correct answer was ‘no’). The mean accuracy rate was 90 % and 93 % and mean reaction times were 436 ms and 513 ms for the learnable and non-learnable categories, respectively, across the whole training session. There was a clear effect of training in the accuracy and reaction time measures with improving performance towards the end of the session as shown in Figure 5. This was evidenced in Category x Quarter ANOVA for accuracy that showed a main effect of Quarter [$F(3,27)=109.449$, $p<0.001$, $\eta^2_p=0.924$] and for the reaction times that showed main effects of Category [F(1,9)=11.028, p<0.010, η^2_p=0.551] (faster reaction time for the learnable than non-learnable category) and Quarter [F(3,27)=23.838, p<0.001, η^2_p=0.726].

3.1.2 MEG results

The MEG data showed clear visual and auditory evoked fields (Figure 6 and 7). In addition, starting at around 350 ms a slowly evolving response was observed whose strength differentiated between the categories after the first quarter of the training.

The largest difference between the categories appeared in the parieto-occipital gradiometers, as a slowly evolving response. The response was similar for the two categories during the first quarter of the session, started to differ between categories during the second quarter, and remained different between categories until the end of the session.

The distributed source analysis paralleled the sensor level trends. Several cortical regions were selected for analysis based on our hypothesis on involvement of the sensory cortices and cross-modal integration areas (Figure 3, see Materials and methods). Activation loci were found in the left and right inferior temporo-occipital areas as well as left frontal areas and right central-parietal areas in the time window of the slowly growing difference between the categories (Figure 8). Three cortical regions showed statistically significant differences in source strength between the categories as a function of training (Figures 9 and 10). The repeated measures ANOVAs showed Category x Quarter interaction for the source activity originating from the temporo-occipital areas at 500 – 650 ms [F(3,36)=4.710, p<0.008, η^2_p=0.282]) as well as the caudal middle frontal cortex at 350 – 500 ms [F(3,36)=8.287, p<0.001, η^2_p=0.408] and 500 – 650 ms [F(3,36)=14.305,

$p<0.001$, $\eta^2_p=0.544$]). Category x Quarter x Hemisphere interaction reached significance in the posterior superior temporal sulcus at 350 – 500 ms [$F(3,36)=5.487$, $p<0.004$, $\eta^2_p=0.314$]). Figure 6 displays the post hoc t-test results for paired comparisons. The temporo-occipital and caudal middle frontal cortices showed an increase in activation from the first quarter of training to the second, third and fourth quarters of the training. The posterior superior temporal sulcus area showed only one statistically significant difference: the first and second quarters had different source strength only in the non-learnable category. This was due to a decrease of source strength from the first to the second quarter.

3.2 Experiment 2: Passive learning

Similarly to the active learning experiment, the EEG data for the passive learning was examined in four blocks of equal length (10 min). The ERP waveforms are shown in Figure 11 and corresponding ERP topographies at three time points in Figure 12. The ERPs in each block showed a statistically significant difference between the two categories (Figure 13) (Block 1, cluster 1: $p <0.035$; Block 2, cluster 1: $p<0.004$, cluster 2: $p<0.019$, cluster 3: $p<0.025$; Block 3, cluster 1: $p<0.009$, cluster 2: $p<0.021$; Block 4, cluster 1: $p<0.001$, cluster 2: $p<0.001$; cluster 3: $p<0.001$; cluster 4: $p<0.005$; cluster 5: $p<0.036$; for details see Table 1). To further examine when the categories start to diverge from each other the first block was divided into two 5-minute sub-blocks and the categories were again compared (Figure 14). There were no statistically significant condition differences ($p = 0.274$) in the ERPs measured during the first 5 minutes whereas the between-category differences during the second 5-minute sub-block were statistically significant (cluster 1, $p < 0.045$ at 165-276 ms, fronto-central distribution) (Figure 15). The differences were most prominent at the frontal and central channels between 200 and 500 ms for each comparison, with the exception that Block 3 showed differences between the categories at an earlier time window.

Table 1. Cluster details of the cluster-based permutation statistics between the responses to the learnable and non-learnable stimuli.

	Cluster mean (μV): Learnable	Cluster mean (μV): Non-learnable	Difference (μV)	Cluster time window (ms)	Cluster location
1st 10 min	-1.55	-0.86	0.69	236-482 ms	R Frontal
2nd 10 min	1.62 / -0.71 / 1.06	1.18 / -0.52 / 0.69	0.44 / 0.19 / 0.37	189-455 ms / 312-563 ms / 95-174 ms	Parietal / L Frontal / L Parietal
3rd 10 min	-0.42 / 0.03	0.14 / 0.34	0.56 / 0.31	172-252 ms / 26-73 ms	R Occipital / L Fronto-central
4th 10 min	1.48 / -1.50 / 0.68 / -0.65 / -1.50	0.74 / -0.54 / 0.26 / -0.04 / -2.08	0.74 / 0.96 / 0.42 / 0.61 / 0.58	328-591 ms / 349-572 ms / 162-293 ms / 172-253 ms / 195-303 ms	Central / L Frontal / Fronto-central / Occipital / R Temporal

Note. Multiple clusters are separated by strokes. R = right, L = left

4. Discussion

Learning of audio-visual associations is an important ability which is crucial for example in reading acquisition. Here we examined the activation dynamics of brain areas utilized at the initial stages of audio-visual association learning. We expected to see learning effects at the early sensory responses as well as in a later time window linked to perceptual learning and audio-visual integration in brain areas that previous studies have linked to short-term cross-modal learning (e.g., Raij et al., 2000; Hashimoto & Sakai, 2004). Indeed, with the progression of the audio-visual association training in MEG and EEG, a gradual increase in brain activity was identified at around 350 ms after stimulus presentation regardless of the learning method (active training or passive exposure). In the passive EEG experiment an earlier effect around 200 ms, corresponding to the auditory P2 response, was additionally observed.

Using MEG, these learning-related changes were localized bilaterally to the caudal middle frontal (CMF) cortex starting at around 350 ms and in the temporo-occipital (TO) area starting at around 500 ms from the visual stimulus onset. These changes in the activity were specific for the stimuli where the audio-visual association could be learned whereas for the control stimuli where no association could be learned the activity remained relatively unchanged throughout the training. Additionally, the right posterior superior temporal sulcus (pSTS) showed modulation in activity during the training at 350 - 500 ms after the visual stimulus onset. All of these learning effects emerged after the first 10 minutes of training. Particularly the TO area has been implicated in cross-modal processing in previous studies and has been linked with short-term learning (e.g., Calvert, 2001; Hashimoto & Sakai, 2004, Blomert, 2011). These learning effects in brain activity coincided with improvement in accuracy scores and reaction times at the group level.

Both TO and pSTS areas have been implicated in reading and the processing of letter-speech sound combinations, with typically an emphasis on the left hemisphere (e.g., van Atteveldt et al., 2009; Raij et al., 2000; Hashimoto & Sakai, 2004; Blomert, 2011). The current results showed bilateral changes in activation for the TO area ~~and right hemispheric modulation of pSTS activity~~ as the training progressed. Further, a change in pSTS activity occurred only for the non-learnable stimuli in the right hemisphere. These differences in the lateralization pattern could reflect the difference in experimental designs between the present and previous studies with examination of learning during the training process in the present study whereas previous studies have used a pre vs. post measurement design. This initial stage of learning may recruit brain areas bilaterally in the case of the TO area~~ or with even right hemispheric preponderance~~. The role of the change in pSTS activity is not clear and is counter to our expectation that activation changes would be observed for the learnable stimuli and not for the non-learnable stimuli.

Frontal cortices also showed enhanced activity bilaterally after 10 minutes of training. The caudal middle frontal cortex most likely reflects either working memory or attentional control in the current experiment (Calvert, 2001; Andersson et al., 2009; Kastner & Ungerleider, 2000; Moisala et al., 2015). Although there is some evidence that the frontal cortices might also be involved in cross-modal association learning (Calvert, 2001; Fuster et al., 2000) and cross-modal working memory tasks (Zhang et al., 2004) it is likely that the frontal activation reflects some form of assistive processing in the form of selective attention. This would also fit into theoretical frameworks of perceptual learning where attentional weighting could manifest as increased attention to important dimensions or features of the training material (Goldstone, 1998). However, the frontal cortex activity should be targeted for more careful study in future experiments in order to disambiguate the cognitive processes.

The spatial pattern of brain activity in the current MEG study was comparable to that found using fMRI by Hashimoto and Sakai (2004) with the exception that, in the current study, the parieto-occipital area did not show strong activation and that the inferior temporo-occipital activity was bilateral and not left lateralized (cf. Fig. 7D in Hashimoto & Sakai, 2004). The bilateral activity observed here might be related to the unfamiliar nature of the rotated Hiragana letters to Finnish speakers who would process them more as complex pictures than as letters, unlike the native Japanese speakers in the study by Hashimoto and Sakai (2004). In addition, the lack of clear activation in the parieto-occipital area could be related to methodological differences. Here the first 600 ms of brain activity were examined after the symbol presentation whereas in fMRI studies activity is integrated over a longer time window. Furthermore, an extended training session might be needed to detect parieto-occipital activity, with possibly even an overnight consolidation period between the measurement points.

Activity in the temporo-occipital and caudal middle frontal areas started to change already after the first 10 minutes of training in both the left and right hemisphere. Interestingly, the differentiation between the trials in the learnable and non-learnable categories started at 350 ms after the visual stimulus onset in the frontal cortex but at 500 ms in the temporo-occipital area. A similar effect of earlier activation of the prefrontal areas than occipital visual areas has been found in an MEG study on declarative memory formation for single presentations of visual stimuli (Takashima et al., 2006). In that study, the effect was interpreted as a top-down interaction to focus on the most important aspects of the shown stimulus (Takashima et al., 2006). However, in the current study, each visual stimulus was simple and repeated altogether 120 times. Therefore, it is unlikely that focusing on certain aspects of the stimulus could account for the current result.

It is also possible that not only selective attention but also working memory processes were reflected in the frontal cortex activity. In this framework, the time window starting at around 350 ms could be interpreted as the retention period in working memory during which sustained neuronal activity has been suggested to reflect active maintenance of the stimulus representations (Jensen & Tesche, 2002). During this retention period, more attention could be directed towards the stimuli that can be learned, and this process would, in turn, enhance activation in the temporo-occipital areas. This latter activation could represent the maintenance of the visual stimulus in working memory for comparison with the auditory stimulus that was presented with a time delay relative to the visual stimulus (cf. Kastner & Ungerleider, 2006; Courtney et al., 1997). However, the EEG experiment suggests that the modulation of brain activity at this time window is largely automatic and could represent attention allocation without full awareness or audio-visual integration processes that are starting to form during the learning. This conclusion is based on the passive nature of the exposure in the EEG experiment and on the simultaneous presentation of the auditory and visual stimuli, which make working memory processes unlikely cause for the observed effects.

The EEG findings supported the conclusion that a cross-modal learning process was taking place at the time window after 300 ms. The time window after 300 ms matches well with the current active learning task and with earlier EEG studies examining audio-visual learning using a congruency manipulation (Shams et al., 2005; Karapidis et al., 2017; 2018). Previously brain activity at this time window has been shown to occur after a relatively short training period that was carried out a few days before the measurement of brain activity (Shams et al., 2005; Karapidis et al., 2017; 2018). However, given the differences in the response topography and experimental design it is likely that the process observed after 300 ms reflects different neuronal mechanisms in the different studies. In previous studies it has been linked to processing of incongruent audio-visual information (Shams et al., 2005; Karapidis et al., 2017; 2018) whereas in our study it is linked to enhanced neural resources for the learning stimuli compared to control stimuli. Regardless of these differences our results show that changes in brain mechanisms related to audio-visual integration at a relatively late time window can be observed already during the training. Due to the passive nature of the EEG experiment, the observed learning process after 300 ms appears rather automatic and occurs without explicit instruction to learn the audio-visual associations.

The current results would be interesting to link with the earlier studies examining the responses to incongruent audio-visual stimuli (e.g., Karapidis et al., 2017; 2018) and perceptual tuning of ambiguous speech stimuli by reading exposure (Bonte et al., 2017). This would require an experiment where the learning effects would be tested across several days to examine if these initial stages of learning would lead to further functional changes observed in the earlier studies. We would predict that the strength of the effects at the initial learning stage would correlate with how rapidly the learnt associations become automatic (reflected in emergence of congruent-incongruent stimulus difference). Further, we would predict that the size of the learning effect at the initial stage would link with the strength of the perceptual tuning for ambiguous phonemes after exposure to written material (Bonte et al., 2017).

Interestingly, the EEG experiment additionally revealed a learning effect at a time window around 200 ms, corresponding to the auditory P2 response, which was not observed for the MEG experiment (in which only the visual responses were examined). The effect emerged rapidly, already after 5 minutes of exposure. The P2 has been linked to attentional processes, auditory object representation as well as sound feature encoding and stimulus classification (Crowley & Colrain, 2004; Ross et al., 2013). It could be that a pre-attentive process to the learnable stimuli was first started which then led to further learning effects at 300 ms and later.

Changes in the auditory P2 have been observed in earlier studies following active training and exposure to auditory stimuli (Reinke et al., 2003; Sheehan et al., 2005; Tremblay et al., 2010; Ross et al., 2013). For example, Sheehan and colleagues (2005) found that the auditory P2 response to repeated sounds increases during the EEG recording, showing larger amplitudes during the second half of the recording than the first half. A similar observation has been made with repeated measurement sessions with auditory stimuli (Ross et al., 2013). Besides replicating these earlier effects of the fast enhancement of the P2 amplitude our finding also shows that, but that the P2 amplitude is modulated by the consistency of the cross-modal stimulus presentation. This suggests that the P2 enhancement is sensitive to and reflecting learning of regularities in the environment. We suggest that this initial sensitivity to the cross-modal regularities in the environment made the learnable stimuli more salient, and this information led to later cross-modal processes that could result in behaviourally relevant changes as found in previous studies (e.g., Karipidis et al., 2017; 2018).

In the current study, brain activity during audio-visual association learning tasks could be tracked using MEG and EEG measures. After 5 to 10 minutes of training to associate auditory and visual material, changes in brain activity were observed in the frontal and temporo-occipital cortices bilaterally at around 350 ms after stimulus presentation, with additional modulation of brain activity around the auditory P2 response. We propose that cross-modal regularities can be extracted already around 200 ms and this can lead to the initial steps in formation of audio-visual associations using top-down selective attention mechanisms.

5. References

Andersson, M., Ystad, M., Lundervold, A., Lundervold, A. J. 2009. Correlations between measures of executive attention and cortical thickness of left posterior middle frontal gyrus-a dichotic listening study. *Behav. Brain Funct.* 5, 41.

Bernstein, L. E., Liebenthal, E. 2014. Neural pathways for visual speech perception. *Front. Neurosci.* 8, 386. 10.3389/fnins.2014.00386.

Blau, V., van Atteveldt, N., Ekkebus, M., Goebel, R., Blomert, L. 2009. Reduced neural integration of letters and speech sounds links phonological and reading deficits in adult dyslexia. *Curr. Biol.* 19, 503-508.

Blomert, L. 2011. The neural signature of orthographic–phonological binding in successful and failing reading development. *NeuroImage.* 57, 695-703.

Bonte, M., Correia, J. M., Keetels, M., Vroomen, J., Formisano, E. 2017. Reading-induced shifts of perceptual speech representations in auditory cortex. *Sci. Rep.* 7, 5143.

Brem, S. et al. 2010. Brain sensitivity to print emerges when children learn letter–speech sound correspondences. *Proc. Nat. Acad. Sci. USA.* 107, 7939-7944.

Calvert, G. A. 2001. Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cer. Cor.* 11, 1110-1123.

Cohen, L. et al. 2000. The visual word form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain.* 123, 291 – 307.

Courtney, S. M., Ungerleider, L. G., Keil, K., Haxby, J. V. 1997. Transient and sustained activity in a distributed neural system for human working memory. *Nature.* 386, 608.

Desikan, R. S. et al. 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage.* 31, 968-980.

Fonov, V. et al. 2011. Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage.* 54, 313-327.

Fuster, J. M., Bodner, M., Kroger, J. K. 2000. Cross-modal and cross-temporal association in neurons of frontal cortex. *Nature.* 405, 347-351.

Garcés, P., López-Sanz, D., Maestú, F., Pereda, E. 2017. Choice of magnetometers and gradiometers after signal space separation. *Sensors.* 17, 2926.

Goldstone, R. L. 1998. Perceptual learning. *Annu. Rev. Psychol.* 49, 585-612.

Gonzalo, D., Shallice, T., Dolan, R. 2000. Time-dependent changes in learning audiovisual associations: a single-trial fMRI study. *NeuroImage.* 11, 243-255.

Hashimoto, R., Sakai, K. L. 2004. Learning letters in adulthood: Direct visualization of cortical plasticity for forming a new link between orthography and phonology. *Neuron.* 42, 311-322.

Jensen, O., Tesche, C. D. 2002. Frontal theta activity in humans increases with memory load in a working memory task. *Eur. J. Neurosci.* 15, 1395–1399.

Karipidis, I. I., Pleisch, G., Röthlisberger, M., Hofstetter, C., Dornbierer, D., Stämpfli, P., Brem, S. 2017. Neural initialization of audiovisual integration in prereaders at varying risk for developmental dyslexia. *Hum. Brain Mapp*. 38, 1038-1055.

Karipidis, I. I., Pleisch, G., Brandeis, D., Roth, A., Röthlisberger, M., Schneebeli, M., ... & Brem, S. 2018. Simulating reading acquisition: The link between reading outcome and multimodal brain signatures of letter–speech sound learning in prereaders. *Sci. Rep.* 8, 7121.

Kastner. S., Ungerleider, L. G. 2000. Mechanisms of visual attention in the human cortex. *Annu. Rev. Neurosci.* 23, 315-341.

Kayser, C., Petkov, C. I., Logothetis, N. K. 2009. Multisensory interactions in primate auditory cortex: fMRI and electrophysiology. *Hear. Res.* 258, 80-88.

Maris, E., Oostenveld, R. 2007. Nonparametric statistical testing of EEG-and MEG-data. *J. Neurosci. Met.* 164, 177-190.

Maurer, U., Zevin, J. D., McCandliss, B. D. 2008. Left-lateralized N170 effects of visual expertise in reading: Evidence from Japanese syllabic and logographic scripts. *J. Cogn. Neurosci.* 20, 1878-1891.

Moisala, M. et al. 2015. Brain activity during divided and selective attention to auditory and visual sentence comprehension tasks. *Front. Human Neurosci.* 9, 86.

Molholm, S. et al. 2006. Audio-Visual Multisensory Integration in Superior Parietal Lobule Revealed by Human Intracranial Recordings. *J. Neurophysiol.* 96, 721-729.

Murray, M. M., Thelen, A., Thut, G., Romei, V., Martuzzi, R., Matusz, P. J. 2015. The multisensory function of the human primary visual cortex. *Neuropsychologia.* 83, 161-169.

Paraskevopoulos, E., Herholz, S. 2013. Multisensory integration and neuroplasticity in the human cerebral cortex. *Transl. Neurosci.* 4, 337-348.

Perrin, F., Pernier, J., Bertrand, O., Echallier, J. F. 1989. Spherical splines for scalp potential and current density mapping. *Electroencephal. Clin. Neurophysiol*. 72, 184-187.

Raij, T., Uutela, K., Hari, R. 2000. Audiovisual integration of letters in the human brain. *Neuron.* 28, 617-625.

Reinke, K. S., He, Y., Wang, C., Alain, C. 2003. Perceptual learning modulates sensory evoked response during vowel segregation. *Cogn. Brain Res.* 17, 781-791.

Ross, B., Jamali, S., Tremblay, K. L. 2013. Plasticity in neuromagnetic cortical responses suggests enhanced auditory object representation. *BMC Neurosci.* 14, 151.

Shams, L., Iwaki, S., Chawla, A., Bhattacharya, J. 2005. Early modulation of visual cortex by sound: an MEG study. *Neurosci. Lett.* 378, 76-81.

Shams, L., Seitz, A. R. 2008. Benefits of multisensory learning. *Trends Cogn. Sci.* 12, 411-417.

Sheehan, K. A., McArthur, G. M., Bishop, D. V. 2005. Is discrimination training necessary to cause changes in the P2 auditory event-related brain potential to speech sounds?. *Cogn. Brain Res.* 25, 547-553.

Tadel, F., Baillet, S., Mosher, J. C., Pantazis, D., Leahy, R. M. 2011. Brainstorm: a user-friendly application for MEG/EEG analysis. *Comput. Intell. Neurosci.* 8, 879716. 10.1155/2011/879716.

Takashima, A., Jensen, O., Oostenveld, R., Maris, E., Van de Coevering, M., Fernandez, G. 2006. Successful declarative memory formation is associated with ongoing activity during encoding in a distributed neocortical network related to working memory: a magnetoencephalography study. *Neuroscience.* 139, 291-297.

Tanabe, H. C., Honda, M., Sadato, N. 2005. Functionally segregated neural substrates for arbitrary audiovisual paired-association learning. *J. Neurosci.* 25, 6409-6418.

Taulu, S., Simola, J., Kajola, M. 2005. Applications of the signal space separation method. *IEEE Trans. Signal Proc.* 53, 3359-3372.

Tremblay, K. L., Inoue, K., McClannahan, K., Ross, B. 2010. Repeated stimulus exposure alters the way sound is encoded in the human brain. *PLoS One.* 5, e10283.

Yoncheva, Y. N., Blau, V. C., Maurer, U., McCandliss, B. D. 2010. Attentional focus during learning impacts N170 ERP responses to an artificial script. *Dev. Neuropsychol.* 35, 423-445.

Younger, J. W., Booth, J. R. 2018. Parietotemporal Stimulation Affects Acquisition of Novel Grapheme-Phoneme Mappings in Adult Readers. *Front. Hum. Neurosci.* 12, 109.

van Atteveldt, N., Formisano, E., Goebel, R., Blomert, L. 2004. Integration of letters and speech sounds in the human brain. *Neuron.* 43, 271-282.

van Atteveldt, N., Roebroeck, A., Goebel, R. 2009. Interaction of speech and script in human auditory cortex: insights from neuro-imaging and effective connectivity. *Hear. Res.* 258, 152-164.

Zhang, D. et al. 2004. Cross-modal temporal order memory for auditory digits and visual locations: An fMRI study. *Hum. Brain Mapp.* 22, 280-289.

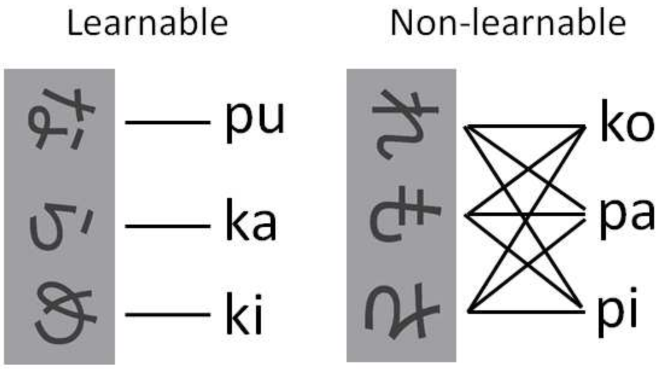


Figure 1. Stimuli in the two categories of trials in the MEG experiment. The lines represent the intended association learning outcomes between auditory (syllables) and visual (symbols) stimuli.

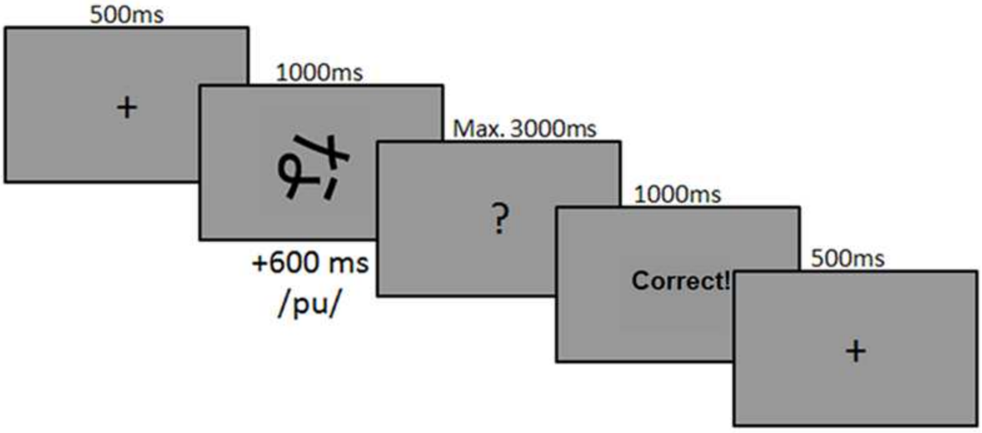


Figure 2. MEG experimental design. A fixation cross appeared at the beginning of each trial, followed by presentation of the visual symbol and, 600 ms later, by the auditory syllable. At 1000 ms after the visual stimulus onset, a question mark appeared to prompt a response. The response was followed by feedback (the word "correct" or "incorrect").

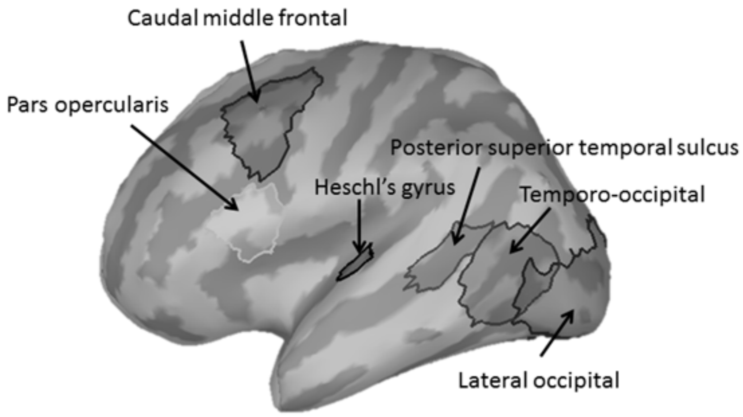


Figure 3. Brain regions of interest. Areas selected for statistical analyses were based on the Desikan-Killiany parcellation and on grand average source-level maps (temporo-occipital area).

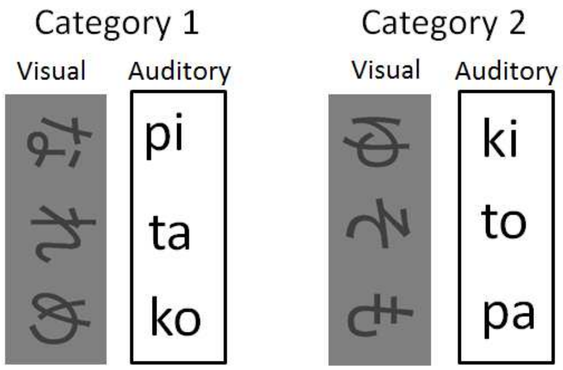


Figure 4. Stimuli in the EEG experiment. Category 1 and 2 items were used as learnable and non-learnable stimulation, counterbalanced across participants.

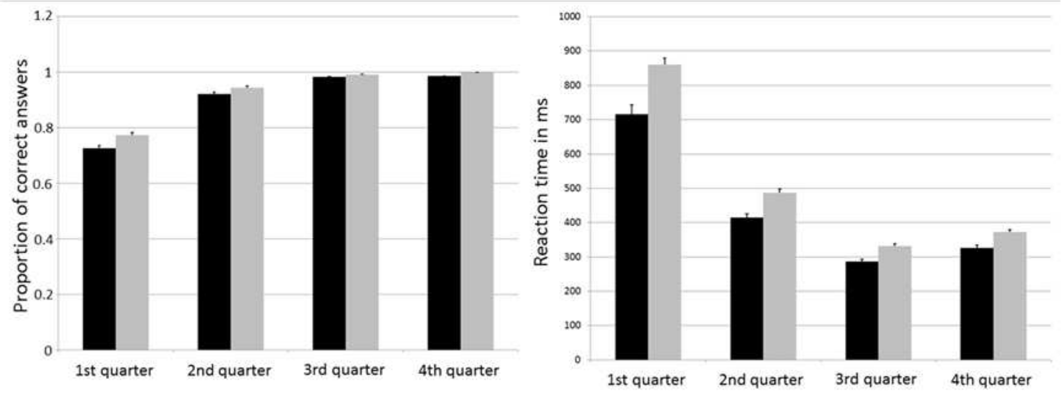


Figure 5. Behavioral effects (N=10). Accuracy (left) and reaction times in milliseconds (right) for the 1st, 2nd, 3rd and 4th quarters of the experiment. Black = learnable, grey = non-learnable. Error bars show the standard error of mean.

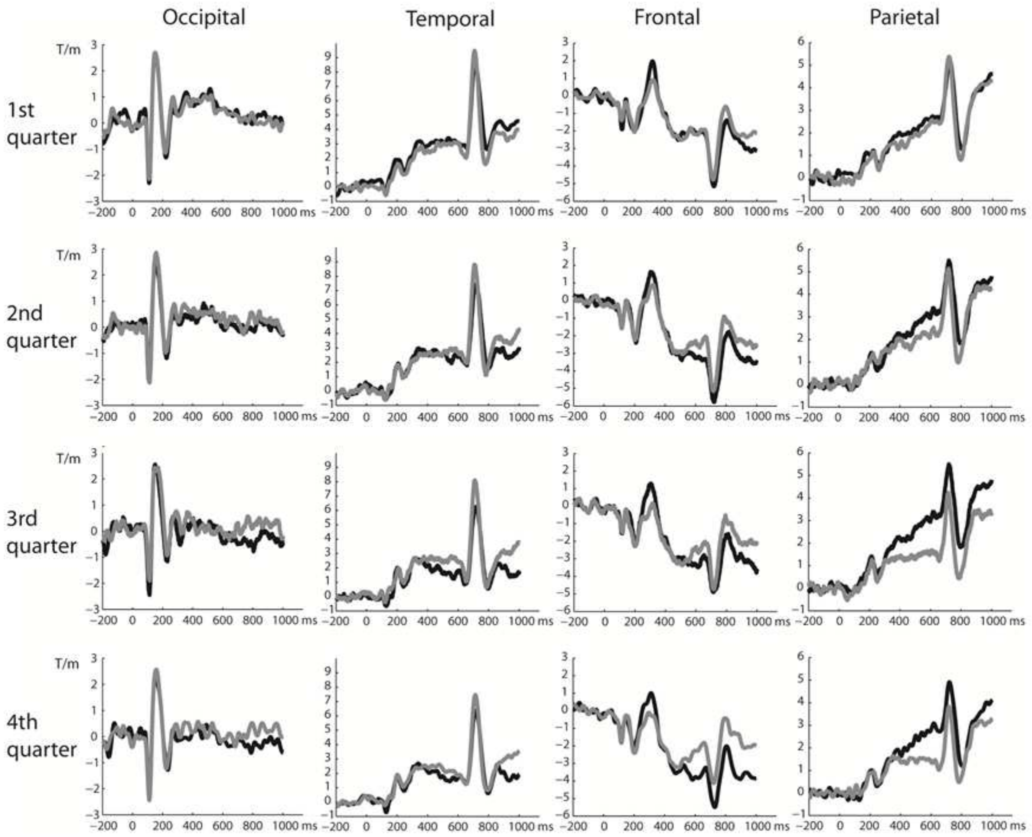


Figure 6. MEG sensor signals (N=13). Selected gradiometers from occipital area representing visual evoked fields, left temporal area representing auditory evoked fields,

as well as left frontal area and left parietal area each representing a slowly evolving evoked field. Black lines = learnable category, grey lines = non-learnable category.

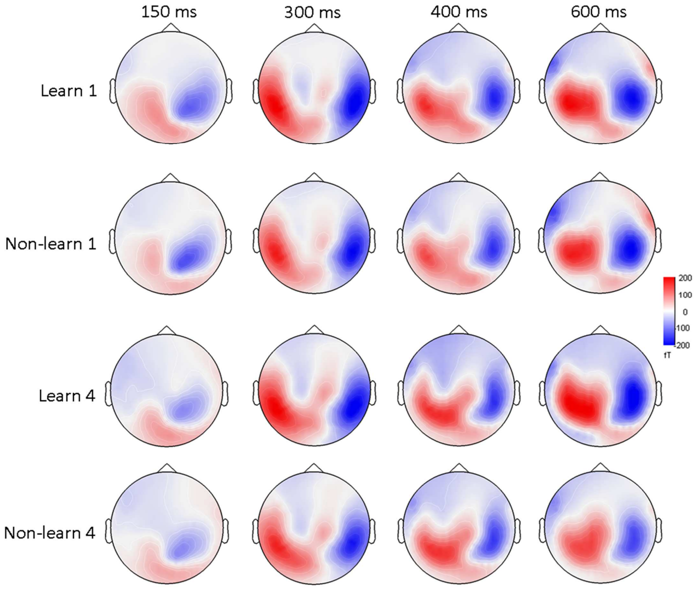


Figure 7. Event-related field topographies at the early visual response time window (150 ms), and later processing time windows (300 ms; 400 ms; 600 ms) before sound onset for the learnable and non-learnable categories during the first 10 minutes and the last 10 minutes of exposure.

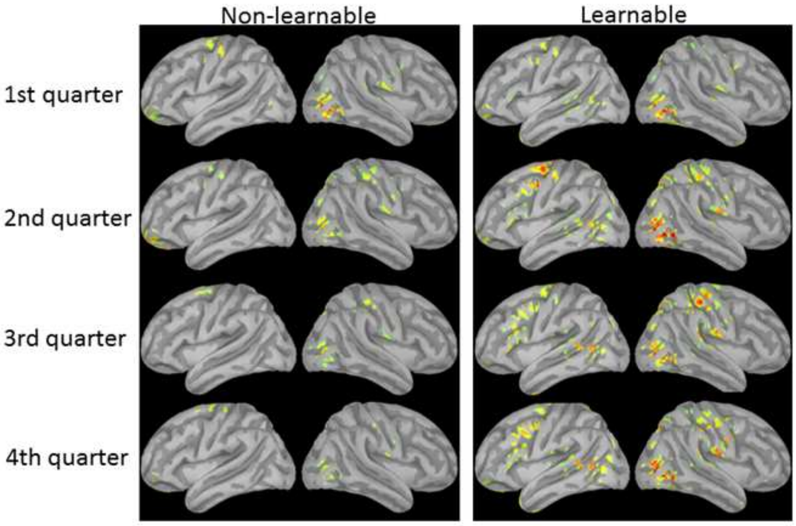


Figure 8. Activated cortical areas (N=13). Grand average source activity (wMNE) during the training at 610 ms after visual symbol onset. Each quarter corresponds to 10 minute block of training.

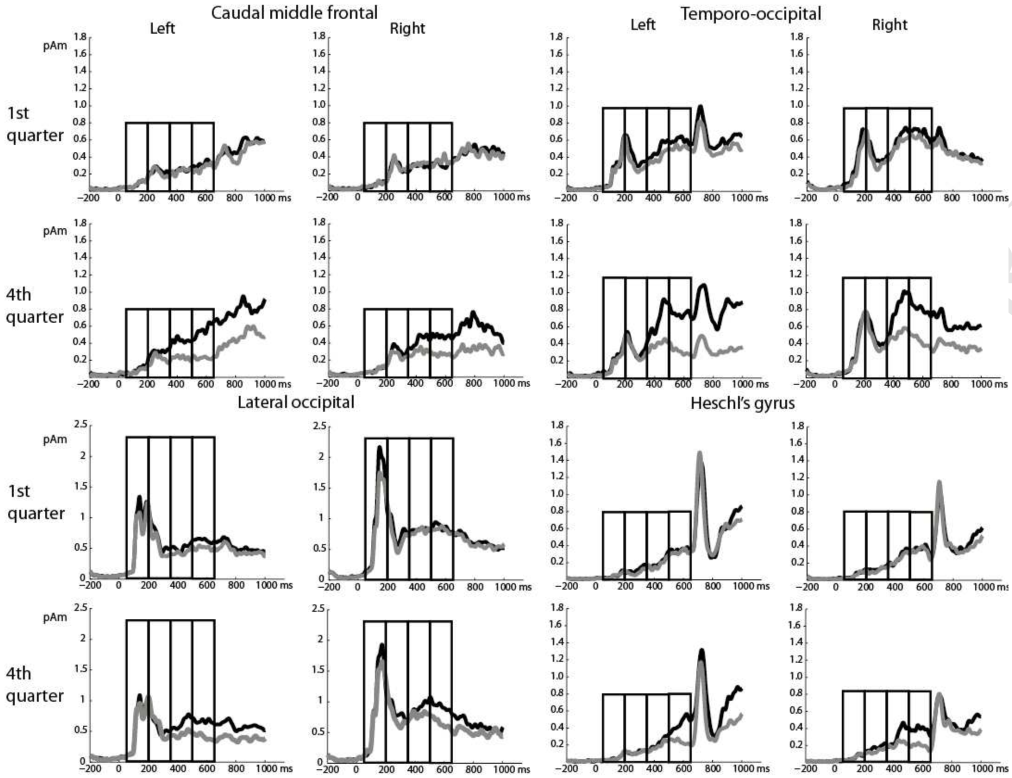


Figure 9. Cortical time course of activation (N=13). Source waveforms in the regions of interests (ROIs), based on wMNE. Black line = learnable category, grey line = non-learnable category. Boxes represent time windows used in the analyses. Note that the amplitude scale for the lateral occipital cortex is different from that for the other ROIs.

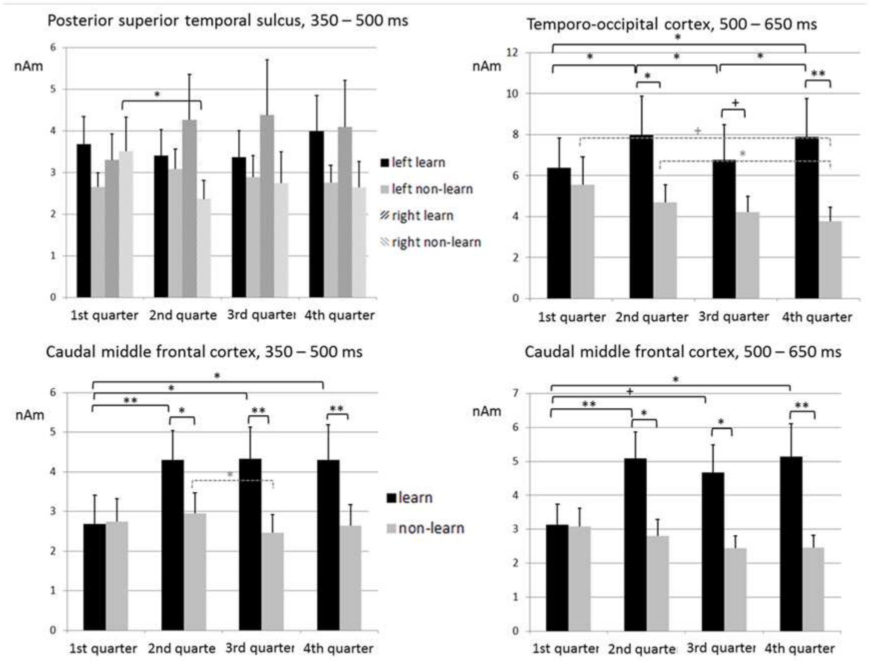


Figure 10. Cortical learning effects (N=13). Source strength values and the significance levels of the post hoc paired t-tests for each quarter of the training for the cortical areas showing training effects in the ANOVA. Note that hemispheres were combined for the temporo-occipital and caudal middle frontal cortices because no hemispheric interaction was found in the ANOVA. Error bars show the standard error of mean. + p<0.10, * p<0.05, ** p<0.010

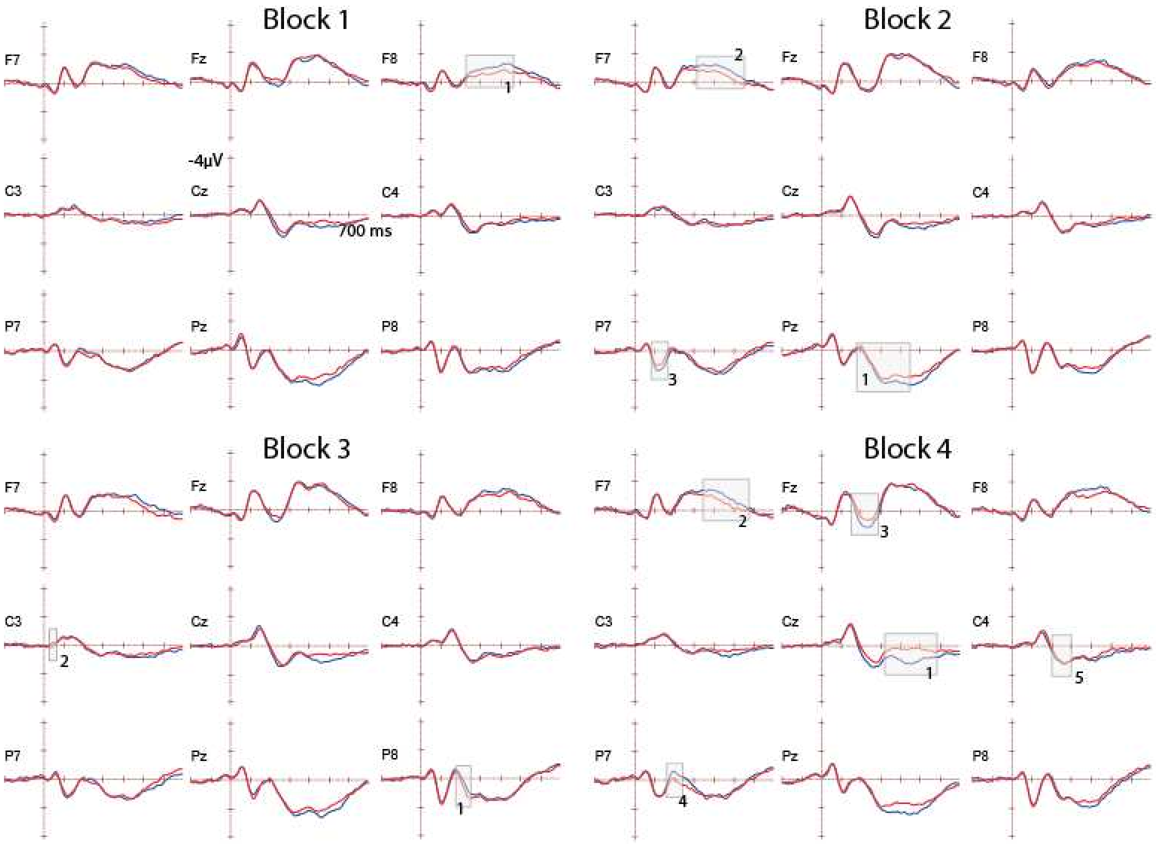


Figure 11. ERP waveforms (N=17) to the audio-visual stimuli presented consistently with the same speech sound and symbol (blue line, learnable category) or inconsistently with the speech-symbol pairs (red line, non-learnable category). Each block represents consecutive 10 minute exposures to the stimulation. The grey box represents the approximate time window for the difference between the stimulus categories given by the cluster-based permutation statistics (the number refers to the cluster number). Vertical marks represent 2 µV, horizontal marks 100 ms, negativity is plotted up.

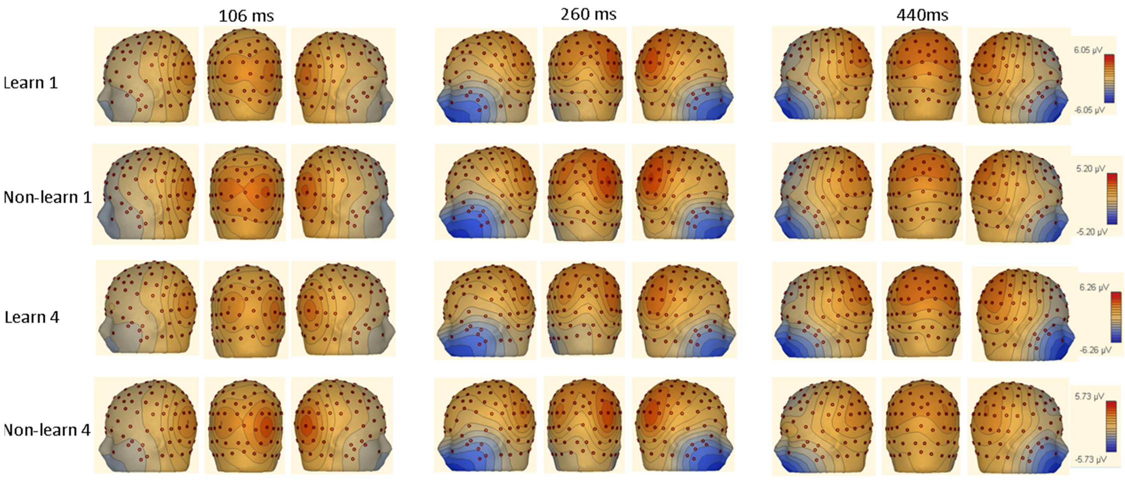


Figure 12. Event-related potential topographies at the P1/N1 time window (106 ms), P2 time window (260 ms) and late time window (440 ms) for the learnable and non-learnable categories during the first 10 minutes and last 10 minutes of exposure.

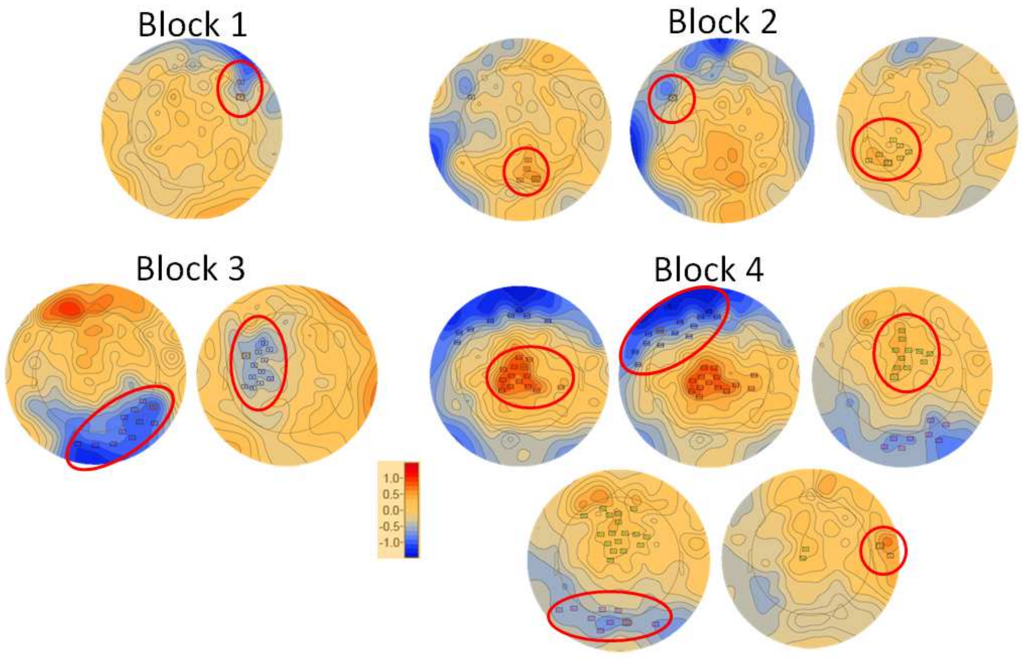


Figure 13. Difference topography (N=17) between the responses to the learnable and non-learnable categories during four consecutive 10-minute blocks of exposure to the audio-visual stimulation. The topographies are displayed at the time when the maximal difference is found between the categories: for Block 1 at 272 ms (cluster 1), for Block 2 at 345 ms (cluster 1), 506 ms (cluster 2) and 117 ms (cluster 3), for Block 3 at 230 ms

(cluster 1) and 45 ms (cluster 2), and for Block 4 at 463 ms (cluster 1), 491 ms (cluster 2), 211 ms (cluster 3), 194 ms (cluster 4) and 262 ms (cluster 5). The rectangles show the EEG channels forming the clusters and the red circles highlight the cluster locations.

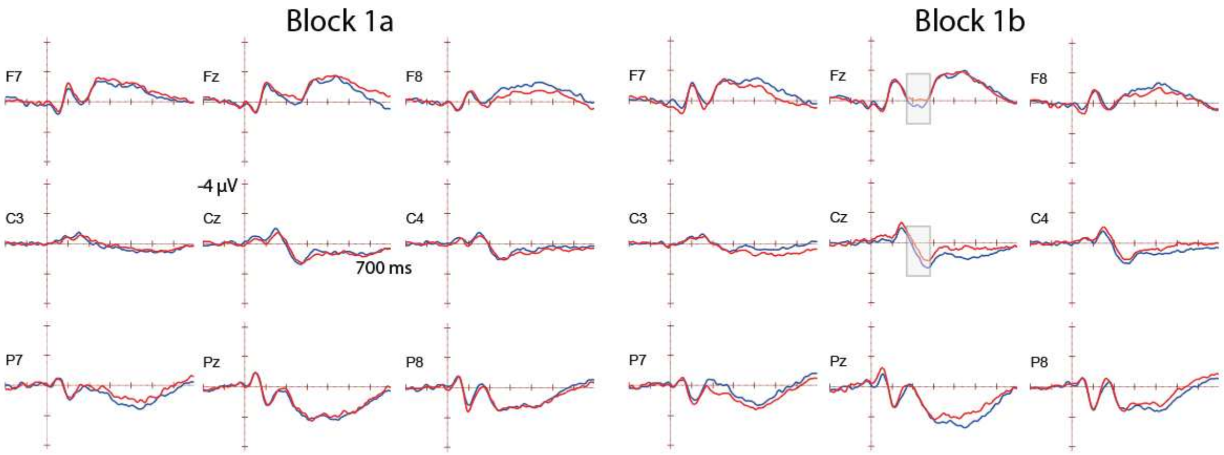


Figure 14. ERP waveforms (N=17) to the first 10 minutes exposure to the audio-visual stimuli presented consistently with the same speech sound and symbol (blue line, learnable category) or inconsistently with the speech-symbol pairs (red line, non-learnable category). Block 1a represents the very first 5 minutes of exposure and Block 1b the second, consecutive 5 minutes of exposure to the stimulation. The grey box represents the approximate time window for the difference between the stimulus categories given by the cluster-based permutation statistics. Vertical marks represent 1 µV, horizontal marks 100 ms, negativity is plotted up.

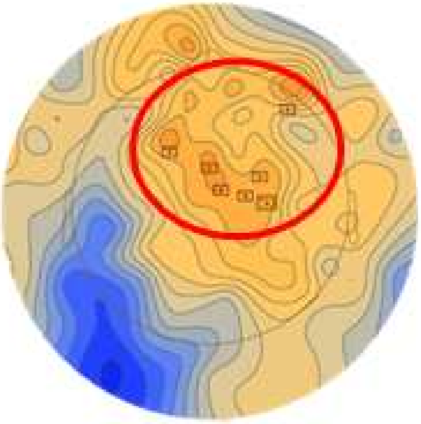


Figure 15. Difference topography (N=17) between the responses to the learnable and non-learnable categories at 205 ms during the second 5 minutes of exposure to the audio-visual stimulation. The rectangles show the EEG channels forming the cluster, and the red circle highlights the cluster location.