35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Semantic-based Sound Retrieval by ERP in Rapid Serial Auditory Presentation Paradigm*

Lei Jiang, Bangyu Cai, Siyuan Xiao, Yiwen Wang*, Weidong Chen and Xiaoxiang Zheng

Abstract—**"Semantic gap" is the major bottleneck of semantic-based multimedia retrieval technique in the field of information retrieval. Studies have shown that robust semantic-based image retrieval can be achieved by single-trial visual evoked event related potential (ERP) detection. However, the question remains whether auditory evoked ERP can be utilized to achieve semantic-based sound retrieval. In this paper, we investigated this question in the rapid serial auditory presentation (RSAP) paradigm. Eight BCI-naïve participants were instructed to perform target detection in RSAP sequences with the vocalizations of 8 familiar animals as sound stimuli, and we compared ERP components and single-trial ERP classification performance between two conditions, the target was a predefined specific one, and the targets were different but belonged to the same semantic category (*i.e.*, semantic-based sound retrieval). Although the amplitudes of ERP components (*e.g.*, N2 and P3) and classification performance decreased a little due to the difficulty of the semantic-based sound retrieval tasks, the best two participants still achieved the area under the receive operating characteristic curve (AUC) of single-trial ERP detection more than 0.77. It suggested that semantic-based sound retrieval by auditory evoked ERP was potentially feasible.**

I. Introduction

The explosively growing of multimedia information in internet era brings challenges to retrieve multimedia objects fast and accurately. Due to the poor ability of text in describing the rich semantic information of multimedia objects, the traditional text-based multimedia object retrieval technique is gradually replaced by semantic-based retrieval technique. Semantic-based retrieval extracts low-level visual and auditory features of the multimedia objects so as to understand the high-level semantic concepts of the objects that matches the retrieval requests [1]. However, due to the "semantic gap" between the low-level visual and auditory feature space that computers can understand and the high-level semantic concept space of humans, it's very difficult for computers to understand the high-level semantic concepts of the multimedia objects like humans [2]. Although people use some methods like ontology inference or machine learning to bridge the semantic gap, the results are still far from satisfaction [3, 4].

An event-related potential (ERP) is a brain response time-locked to an event [5]. Previous studies have reported that some components of ERP (*e.g.*, N400 and P3) are related to semantic search and target detection [6, 7]. Some researchers take advantage of this characteristic of ERP, and try to use human intelligence (some ERP components time-locked to an event onset) in the implementation of semantic-based rapid retrieval of multimedia objects. For example, Paul Sajda and his colleagues tried to combine electroencephalogram (EEG) and technology of computer vision to retrieve images belonging to a specific semantic category [8]. However, semantic-based retrieval of sound, which is another important component of multimedia, has not been fully explored and is still under investigation. Some work shows that brain-computer interfaces (BCIs) based on auditory evoked ERP are not as robust as visual ERP-BCIs [9], which suggests that it would be more difficult to retrieve sounds than images based on ERP.

We are interested in the question "Is it possible to achieve robust semantic-based sound retrieval by auditory evoked ERP?" In the current study, we investigated this question in the rapid serial auditory presentation (RSAP) paradigm. Eight BCI-naïve participants were instructed to perform target detection by button clicking in RSAP sequences, and the EEG signals were simultaneously collected. The potential feasibility of semantic-based sound retrieval based on auditory evoked ERP was explored by comparing ERP components between two conditions, the target was a predefined specific one, and the targets were not specific but belonged to a certain semantic category (*i.e.*, semantic-based sound retrieval). We also used the linear discriminant analysis (LDA) based AdaBoost classifier to detect single-trial ERP in both conditions, and used area under receive operating characteristic curve (AUC) to evaluate the classification performance [10, 11].

II. Methods

A. Experiment Design and Data Acquisition

Eight BCI-naive participants (all males, aged 22-24, all right-handed) participated in this study. No participant had a history of psychiatric or neurological illnesses, and all reported normal hearing. All participants gave written informed consent.

*This work is supported by grants from National High Tech R&D Program of China (No. 2012AA011602), the National Basic Research Program of China (No. 2013CB329506), the National Natural Science Foundation of China (No. 61031002, 61001172, 61233015), and the Fundamental Research Funds for the Central Universities. All the authors are with Qiushi Academy for Advanced Studies, Zhejiang University.

Lei Jiang, Siyuan Xiao and Weidong Chen are also with College of Computer Science and Technology, Zhejiang University, China (e-mail: fishjianglei@gmail.com, yihan2008026@126.com, chenwd@zju.edu.cn). Bangyu Cai and Xiaoxiang Zheng are also with College of Biomedical Engineering & Instrument Science, Zhejiang University, China (e-mail: cbangyu@gmail.com, zxx667@gmail.com).

Yiwen Wang and Xiaoxiang Zheng are also with Key Laboratory of Biomedical Engineering of Ministry of Education, Zhejiang University (phone: 86-571-87952339; fax: 86-571-87952865; e-mail: eewangyw@zju.edu.cn).

The sound stimuli included 120 different vocalizations of 8 familiar animals (frog, fish, dog, tiger, horse, cat, bird, sheep; 15 different exemplars for each animal). Preliminary experiment verified that these vocalizations could be recognized in a very short time (636±24 ms after target stimulus onset) with very high accuracy (90.5±1.3 percent hit rate in target detection). All of these 120 vocalizations were then modified so that they were 500 ms in duration, digitized at 22,050 Hz, 16-bit stereo and saved as WAV format. 10-ms rise/fall times were included to minimize clicks at sound onset and offset. Finally, all vocalizations were normalized.

Participants were required to minimize eye and body movements, and were asked to perform target detection tasks in the RSAP paradigm. Sound stimuli were presented at a rate of 1 Hz with a constant inter-stimulus interval (ISI) of 500 ms through an insert earphone. One of the eight animals mentioned above was randomly selected as the target in a block (*e.g.* 'During this block, press the button to frog croak'). Participants were instructed to press left mouse button when a target (*e.g.* 'frog croak') was detected. Each block consisted of 10 consecutive random sequences of stimuli, *i.e.* 10 trials. In each trial, a vocalization of each animal was presented once respectively. To ensure the quality of ERP, vocalizations belonging to the same animal did not occur in a row between two consecutive trials. Therefore, a block was made of 80 sound stimuli, in which the 10 sound stimuli of each nontarget animal differed from one another which were randomly chosen from the 15 exemplars, while the composition of the 10 sound stimuli of target animal varied in 2 conditions:

1) *Condition I,* "semantic-based target detection". Just like the nontarget stimuli, the 10 sound stimuli of target animal differed from one another which were randomly chosen from the 15 exemplars of target animal. In other words, the target stimuli belonged to the same semantic category (the target animal), but varied in a block;

2) *Condition II,* "specific target detection". One exemplar was randomly chosen from the 15 exemplars of the target animal and served as target stimulus. So target stimuli were constant in a block (*i.e.*, the same target stimulus would be repeated 10 times in a block). Prior to a block, the randomly chosen target stimulus was presented to the participant so as to make him/her clear what the specific target was.

The whole experiment was divided into 4 sessions (two per condition). The order of the sessions was counterbalanced across participants. Each animal served as the target once in a session, in randomized order both within and between participants, making the experiment composed of 32 blocks in total. Prior to the experiment, participants were made familiar with all the sound stimuli materials and the task. Breaks were encouraged between sessions to minimize fatigue and eye movements. The entire experiment lasted about 2.5 h.

The EEG of all participants while performing the tasks were recorded by 60 Ag/AgCl electrodes (impedances < 30 kΩ) at a sample rate of 1000 Hz, referenced to the nose, with a 200 Hz low-pass filter and 50 Hz notch using Neuroscan Synamps system. Electrode positions included the standard 10-20 system locations and intermediate positions. Blinks were monitored with vertical electrooculogram (EOG) recorded from electrodes located above and below the left eye. Electrode AFz served as grounding electrode.

B. Data Analysis and Single-Trial EPR Detection

Button-press responses falling between 300 and 1000 ms post target stimuli onset were considered correct. Blink artifacts were removed from EEG data by correlation between EEG and EOG. Then the EEG data were band-pass filtered from 0.5 to 30 Hz using a second order Butterworth filter. For further analysis, EEG data were epoched from 200 ms before to 800 ms after stimuli onset, with the average of the first 200 ms as baseline. The epochs corresponding to incorrect judgments were excluded from ERP analysis. And the epochs contaminated by excessive eye movements or other artifacts were also discarded as amplitudes exceeding ±80 μv in any EEG channel. Accepted EEG epochs were averaged according to condition (semantic-based target detection, specific target detection) and stimulus type (target stimulus, nontarget stimulus) from each participant to compute the ERP. For identification of ERP components and display purposes, grand group-average ERPs for each of the conditions and stimulus types were also computed. It should be noted that no epoch was rejected in classification analysis, the above-mentioned rejection criteria of EEG epochs was only applied in ERP analysis.

To define the auditory evoked ERP components which were significantly related to attention shift and target detection, firstly, the averaged nontarget epoch was subtracted from the averaged target epoch to get the difference wave for each participant and condition. Then, we performed point-wise running t-tests (two-tailed) to compare the amplitudes of the difference waves to zero at each electrode for each condition. Significance effects of stimulus type (target, nontarget) were defined as at least 40 consecutive data points reaching the 0.05 significance level (40 data points = 40 ms at a 1000 Hz sample rate) [12]. According to the statistical significance of the difference waves, we identified all auditory evoked ERP components significantly related to attention shift and target detection and their time windows. Mean amplitudes of the difference waves of the grand group-average ERPs were computed across these time windows for each electrode to define the scalp topographies of the ERP components. For further statistical analysis, we inspected the scalp topographies of the difference waves to find the scalp regions of maximal mean amplitudes during corresponding time windows. After that, we carried out the statistical analysis on integrated amplitude measurements of the difference waves averaged across the electrodes located on these scalp regions and corresponding time windows.

For offline classification analysis, we used AdaBoost technique based on the LDA classifier to detect single-trial ERP. We used stepwise LDA (SWLDA) to select features for each electrode [13]. We conducted a 5-fold cross-validation for each participant and each condition, and used AUC to evaluate the classification performance.

III. Results and Discussion

In this study, we analyzed the auditory evoked ERPs when participants performed semantic-based sound retrieval tasks

and explored whether the auditory evoked ERPs could be used to achieve semantic-based sound retrieval.

Figure 1 shows the spatiotemporal presentations of the amplitudes of the grand-average difference waves for the two conditions. Electrodes are on the y-axis and time (ms) is on the x-axis. Electrodes are organized in the following order: left hemisphere (LH), midline electrodes, right hemisphere (RH).The time interval selected for analysis was from 0 to 800 ms after stimulus onset. Figure 2 shows significant p-values from point-wise running t-tests for the difference waves. It could be seen that the difference wave of each condition included three prominent components: N1, N2 and P3. For "semantic-based target detection" condition, the time windows of the three components of the difference wave were shown in Figure 2: 0-150 ms, 150-330 ms, 350-700 ms. While for "specific target detection" condition, the time windows were: 0-150 ms, 150-300 ms, 300-670 ms. Figure 3 shows the scalp topographies of these components. It can be seen that the scalp topographies of the two conditions were very similar in morphology, so the statistical analysis of the difference waves was carried on the same scalp regions for the two conditions.

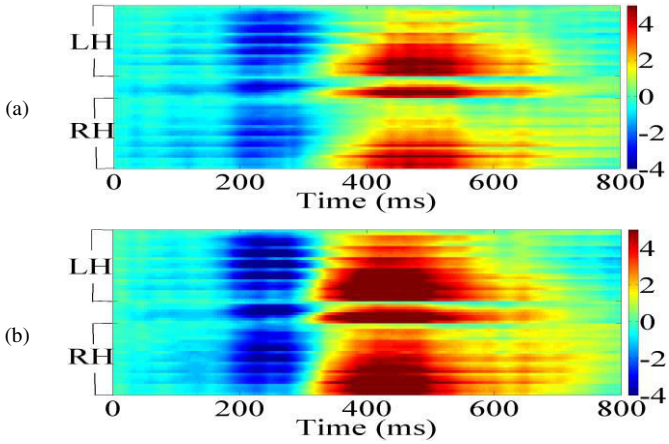

Figure 1. Spatiotemporal presentations of the amplitudes (μv) of the grand-average difference waves for the two conditions. (a) Semantic-based target detection. (b) Specific target detection.

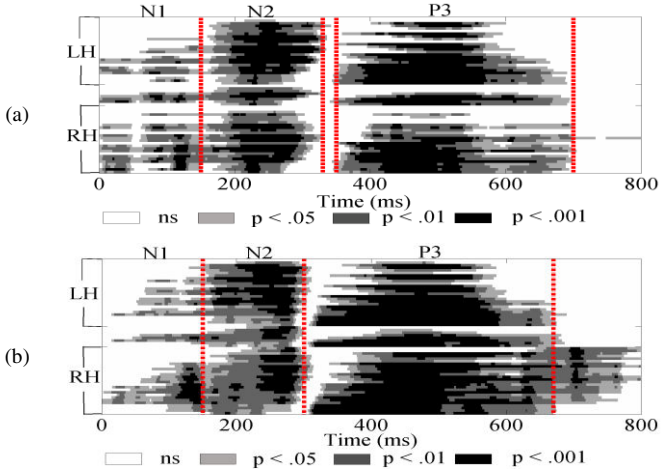

Figure 2. Spatiotemporal presentations of significant p-values from point-wise running t-tests for the difference waves of the two conditions. The time windows of N1, N2, P3 are flagged. (a) Semantic-based target detection. (b) Specific target detection.

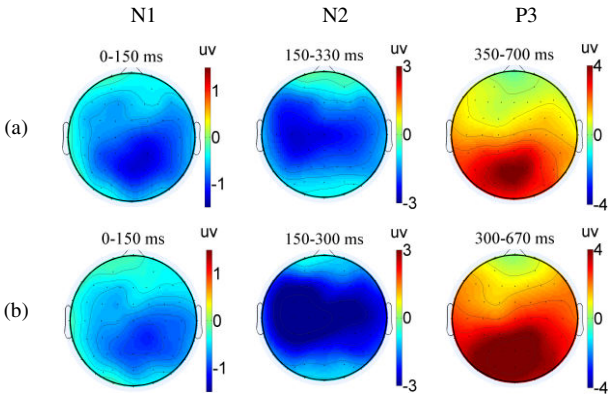

Figure 3. Scalp topographies of the grand-average difference waves of the two conditions. (a) Semantic-based target detection. (b) Specific target detection.

The N1 appeared maximal over centro-parietal (CP1, CPz, CP2), and parietal (P1, Pz, P2) sites, two-tailed paired t-test revealed there was no significant difference between the two conditions for the average amplitudes across these six electrodes (CP1, CPz, CP2, P1, Pz, P2) and corresponding N1 time windows (mean = -1.17 μv vs. mean = -0.94 μv for "semantic-based target detection" versus "specific target detection"; $t(7) = -1.11$, $p > 0.3$). The N2 appeared maximal over fronto-central (FC3, FC1, FCz), central (C3, C1, Cz) and centro-parietal (CP3, CP1, CPz) sites, one-tailed paired t-test revealed that the average amplitudes across these nine electrodes (FC3, FC1, FCz, C3, C1, Cz, CP3, CP1, CPz) and corresponding N2 time windows were greater for "specific target detection" condition (mean = -3.39 μv) than for "semantic-based target detection" condition (mean = -2.31 μv; $t(7) = 2.68$, $p < 0.05$).The P3 appeared maximal over parietal (P1, Pz, P2) and parieto-occipital (PO3, POz, PO4) sites, one-tailed paired t-test revealed that the average amplitudes across these six electrodes (P1, Pz, P2 PO3, POz, PO4) and corresponding P3 time windows were greater for "specific target detection" condition (mean = 4.62 μv) than for "semantic-based target detection" condition (mean = 3.63 μv; $t(7) = -4.81$, $p < 0.001$).

In above-mentioned ERP analysis, the high similarity of scalp topographies between the two conditions revealed that whether the targets were specific or not, to some extent, the activated brain regions may be the same in sound retrieval tasks. The N1 amplitudes did not differ significantly between conditions. Previous studies had reported that auditory evoked N1 was related to the allocation of attention resources for early auditory processing such as stimulus detection and feature encoding [14]. Therefore, we speculated that the attention resources employed in stimulus onset detection and stimulus encoding were unrelated to whether the targets were specific in the current experiment. The N2 amplitudes were significantly different between conditions, with higher amplitudes when the targets were specific. Previous studies had reported that there was a positive correlation between N2 amplitudes and nontarget-target variation [15]. When the targets were specific, the sound retrieval might be more targeted for participants. This may make the targets more distinct from nontargets and then evoke greater N2. The same as N2, the P3 amplitudes were significantly different between

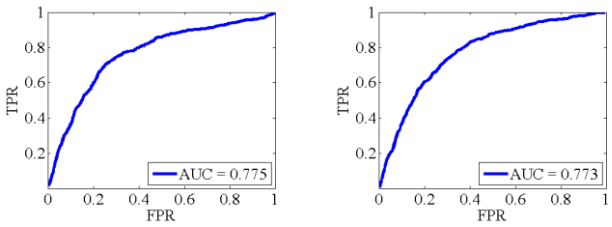

Figure 4. ROC curves of two participants in the "semantic-based target detection" condition.

conditions, with higher amplitudes when the targets were specific. Some published articles reported that more confidence in target recognition was associated with greater P3 amplitude [16]. Participants may be more confident in target recognition when the targets were specific in the sound retrieval tasks and this may be the reason why greater P3 amplitudes in "specific target detection" condition compared with "semantic-based target detection" condition.

However, it was remarkable that two participants reached relatively high classification performance in "semantic-based target detection" condition with AUC >= 0.77 (Figure 4, the x-axis is false positive rate (FPR) and the y-axis is true positive rate (TPR)). In line with the weakening of N2 and P3, the classification performance in the "semantic-based target detection" condition decreased compared with the "specific target detection" condition (0.704±0.055 vs. 0.765±0.069 in average for the eight BCI-naïve participants). The classification performance was lower than the state of the art of rapid image search based on rapid serial visual presentation (RSVP) paradigm [17]. It may because that auditory evoked ERPs were not as robust as visual evoked ERPs in BCIs[9], and the participants in this study were selected randomly and had no BCI experience before. The average classification performance was over 0.70 in the "semantic-based target detection" condition, which indicated the possibility to achieve semantic-based sound retrieval relatively efficiently by single-trial ERP detection [18].

IV. Conclusions

We are interested in whether semantic-based sound retrieval, which is a big challenge in information retrieval, can be achieved by auditory evoked ERP. In this paper, we investigated the amplitudes of ERP components and classification performance when participants performed semantic-based sound retrieval in RSAP tasks, compared with the condition in which the retrieval targets were specific. The amplitudes of ERP components (*e.g.*, N2 and P3) reduced, which resulted in the decrease of the classification performance as expected. However, N2 and P3 were still statistically significant over the baseline, and the AUC in semantic-based sound retrieval task was over 0.70 in average for eight BCI-naive participants. Moreover, the best two participants achieved relatively high performance with AUC > 0.77. The results indicated that semantic-based sound retrieval by auditory evoked ERP was potentially feasible. Future work should explore whether the natural sounds without normalization could be retrieved based on ERP detection. Furthermore, the advance in signal processing techniques and pattern recognition algorithms may contribute to better performance on semantic-based sound retrieval by auditory evoked ERP.

References

[1] M. R. Naphade and T. S. Huang, "Extracting semantics from audio-visual content: the final frontier in multimedia retrieval," *Neural Networks, IEEE Transactions on,* vol. 13, pp. 793-810, 2002.

[2] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP),* vol. 2, pp. 1-19, 2006.

[3] R. Möller and B. Neumann, "Ontology-based reasoning techniques for multimedia interpretation and retrieval," *Semantic Multimedia and Ontologies,* pp. 55-98, 2008.

[4] M. Wang and X. S. Hua, "Active learning in multimedia annotation and retrieval: A survey," *ACM Transactions on Intelligent Systems and Technology (TIST),* vol. 2, pp. 1-21, 2011.

[5] S. J. Luck, *An introduction to the event-related potential technique*, Cambridge, MA: MIT, 2005.

[6] A. Mecklinger, A. F. Kramer, and D. L. Strayer, "Event Related Potentials and EEG Components in a Semantic Memory Search Task," *Psychophysiology,* vol. 29, pp. 104-119, 1992.

[7] S. Politzer-Ahles, R. Fiorentino, X. Jiang, and X. Zhou, "Distinct neural correlates for pragmatic and semantic meaning processing: An event-related potential investigation of scalar implicature processing using picture-sentence verification," *Brain Research,* vol.1490, pp. 134-152, 2012.

[8] E. A. Pohlmeyer, J. Wang, D. C. Jangraw, B. Lou, S. F. Chang, and P. Sajda, "Closing the loop in cortically-coupled computer vision: a brain-computer interface for searching image databases," *Journal of Neural Engineering,* vol. 8, p. 036025, 2011.

[9] A. Furdea, S. Halder, D. Krusienski, D. Bross, F. Nijboer, N. Birbaumer, *et al.*, "An auditory oddball (P300) spelling system for brain-computer interfaces," *Psychophysiology,* vol. 46, pp. 617-625, 2009.

[10] Y. Huang, D. Erdogmus, S. Mathan, and M. Pavel, "Boosting linear logistic regression for single trial ERP detection in rapid serial visual presentation tasks," in *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*, 2006, pp. 3369-3372.

[11] T. Fawcett, "ROC graphs: Notes and practical considerations for researchers," *Machine Learning,* vol. 31, pp. 1-38, 2004.

[12] M. E. Thurlings, A. M. Brouwer, J. B. F. Van Erp, B. Blankertz, and P. J. Werkhoven, "Does bimodal stimulus presentation increase ERP components usable in BCIs?," *Journal of Neural Engineering,* vol. 9, p. 045005, 2012.

[13] D. J. Krusienski, E. W. Sellers, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, "Toward enhanced P300 speller performance," *Journal of Neuroscience Methods,* vol. 167, pp. 15-21, 2008.

[14] R. Näätänen, T. Kujala, and I. Winkler, "Auditory processing that leads to conscious perception: A unique window to central auditory processing opened by the mismatch negativity and related responses," *Psychophysiology,* vol. 48, pp. 4-22, 2011.

[15] S. H. Patel and P. N. Azzam, "Characterization of N200 and P300: selected studies of the event-related potential," *International Journal of Medical Sciences,* vol. 2, pp. 147-154, 2005.

[16] J. Polich, "Updating P300: An integrative theory of P3a and P3b," *Clinical Neurophysiology,* vol. 118, pp. 2128-2148, 2007.

[17] Y. Huang, D. Erdogmus, M. Pavel, S. Mathan, and K. E. Hild, "A framework for rapid visual image search using single-trial brain evoked responses," *Neurocomputing,* vol. 74, pp. 2041-2051, 2011.

[18] H. Cecotti, R. W. Kasper, J. C. Elliott, M. P. Eckstein, and B. Giesbrecht, "Multimodal target detection using single trial evoked EEG responses in single and dual-tasks," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, 2011, pp. 6311-6314.