# Optimizing Visual-to-Auditory Delay for Multimodal BCI Speller

Xingwei An, Dong Ming, Douglas Sterling, Hongzhi Qi, and Benjamin Blankertz

*Abstract*—**Multimodal spellers combining visual and auditory stimulation have recently gained more attention in ERP-based Brain-Computer Interfaces (BCIs). Most studies found an improved efficiency compared to unimodal paradigms while few have explored the effect of the visual-to-auditory delays on the spelling performance. Here, we study five conditions with different visual-to-auditory delays, in order to find the paradigm that provides the best overall BCI performance. We compared the temporal and spatial binary classification accuracy as well as the grand-averaged classification accuracies over repetitions. Results show that long delays may cause better performance in early time intervals corresponding to negative ERP components, but better overall performance is achieved with short visual-to-auditory delays.**

## I. Introduction

Event-related potentials (ERPs) allow to investigate electrical brain activity at a high temporal resolution, ranging from sensory (early ERP components) to higher cognitive processes (later ERP components) [1]. The detection of attention modulated ERPs in single-trials (or averages of few repetitions) are instrumental in one important category of brain-computer interfaces (BCIs), which provide direct and non-muscular communication methods for people with severe motor impairments [2,3]. BCIs that are based on ERPs (such as mental typewriting) are commonly considered to be more stable [4] and more efficient for selection tasks than other paradigms.

Previous work has successfully allowed participants to spell by concentrating on visual and (or) auditory stimuli corresponding to letters. Klobassa et al. [5] used a multimodal audio-visual speller paradigm to provide a better 'training' in initial sessions but using an auditory-only speller in the final test sessions. They used simultaneous visual and auditory stimuli with a presentation time of 110 ms and an inter-stimulus interval of 500 ms. Boll and Berti [6] found that the reaction time prolongation and component amplitudes did not differ significantly between auditory and bi-modal deviants. However, in that study the visual stimulus was presented 80 ms prior to the auditory stimulus to account for the faster transduction process in the inner ear compared to the retina, based on the theory that the visual-to-auditory onset delay allows for visual-auditory interaction [7,8], Senkowski [8] studied audio-visual stimuli that were presented with stimulus onset asynchronies (SOAs) ranging from −125 to +125 ms. The visual stimuli were white horizontal gratings presented on a black background while the auditory stimuli were 1600Hz sinusoidal tones presented at 65 dB.

Recent work has explored using both visual and auditory stimuli ‘simultaneously’ to increase the response in the brain in hope of improving the BCI’s overall accuracy. While these studies have shown such bimodal spellers to be effective, it is not clear how to temporally stagger the auditory and visual stimuli best to maximize the overlapping information conducive to spelling accuracy. Here we discuss an experiment that compares the efficacy of using five different visual-auditory delays. Temporal and spatial binary classification accuracies were compared to find the time interval and channels which enabled the most effective classification.

## II. Material and Methods

### A. Participants

Eleven healthy subjects (3 female) aged 22-34 participated in this study. Two of the participants had already participated in earlier BCI experiments. Each participant provided written informed consent, and did not suffer from a neurological disease and had normal hearing. All participants were included in offline analysis. Participants were all volunteers and were not paid for their participation.

### B. Stimuli

This study compares five different conditions related to sensory modalities that can be used to drive a BCI speller. All of the conditions use two sensory modalities: visual and auditory stimuli. The stimuli are similar to those in existing visual and auditory spellers and each encodes the same information for each selection. The only variable over conditions is the delay between the presentation of corresponding auditory and visual stimuli. All of these paradigms are designed to allow spelling of 30 symbols: the 26 letters of the alphabet, a period ‘.’, a comma ‘,’, a space symbol ‘_’ and a backspace symbol ‘<’ that could be used to erase the previous symbol.

The auditory stimuli were presented through a comfortably positioned light neckband headphone (Sennheiser PMX 200). We choose six stimuli from a similar study [9] which used short spoken syllables (‘it’, ’ti’, ‘to’, ’ot’’) uttered by three speakers (bass, tenor and soprano) as stimuli. These six stimuli were: bass spoken ‘ti’ and ‘to’ (on the left channel), tenor spoken ‘it’ and ‘ot’ (on both channel), and soprano spoken ‘ti’ and ‘to’ (on the right channel). The duration of each stimulus was 130 ms, and we used a stimulus onset asynchrony (SOA) of 200 ms.

Research supported by the China Scholarship Council, National Natural Science Foundation of China (Grant No. 81222021) and the BMBF Grant No. 16SV5839.

Xingwei An, Dong Ming and Hongzhi Qi are with Tianjin University, Tianjin, China. (Corresponding author, e-mail: xing-wei.an@hotmail.com and richardming@tju.edu.cn).

Douglas Sterling (douglas.sterling@bccn-berlin.de) is with the Bernstein Center for Computational Neuroscience in Berlin.

Benjamin Blankertz (benjamin.blankertz@tu-berlin.de) is with Berlin Institute of Technology, Neurotechnology Group, Germany

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

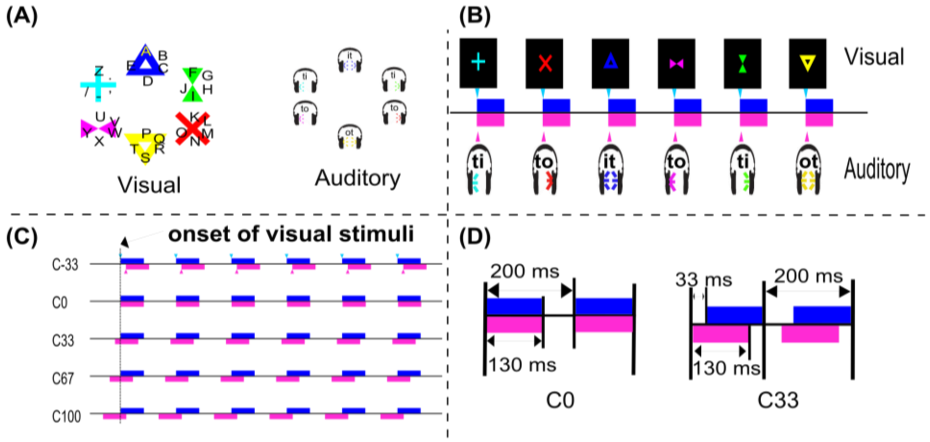

Figure 1. Visualization of the experimental design.

The visual stimuli were presented as in the Center Speller [10], using flashings of 6 different visual shapes, each with a different color (Fig 1.A). These shapes were presented at the center of a 19’ TFT screen with a refresh rate of 60 Hz in a pseudo-random sequence. The presentation of each stimulus lasted 130 ms with a stimulus onset asynchrony (SOA) of 200 ms, chosen to match the duration of the auditory stimuli. Visual stimuli presentation was time-locked to the screen refresh rate. The visual and auditory stimuli were always combined correspondingly in all conditions. Such as, auditory ‘it’ was always combined with visual ‘blue triangle’. Auditory left channel ‘ti’ was always combined with visual ‘green hourglasses.

In all conditions, the selections of each symbol are coded into two step selections (first selection for group, second selection for symbol) of one out of six stimuli, cf. [10]. Conditions are named as ‘C-33’, ’C0’, ’C33’, ’C67’ and ‘C100’. In condition ‘C0’, the timing of the stimulus presentation was set such that visual stimulus was presented simultaneously with the auditory stimulus (See Fig 1.B). The auditory stimuli in other conditions are presented at different frame times relative to the visual stimuli. For instance, in the ‘C-33’ condition auditory stimuli were presented 2 frames (about 33 ms) after the corresponding visual stimuli. However, conditions ‘C33’, ’C67’, and ‘C100’ were designed such that auditory stimuli were presented 2, 4 and 6 frames, respectively, before the corresponding visual stimuli. Fig 1.C shows the time sequences of five conditions. Fig 1.B takes condition ‘C0’ as an example. Blue block stands for the duration of visual stimuli, while the pink one stands for auditory stimuli. Fig 1.D shows the delays of two conditions “C0” and ‘C33’ for a presentation of single visual and its corresponding auditory stimulus.

### C. *Procedure*

Electroencephalogram (EEG) signals were acquired using a Fast’n Easy Cap (EasyCap GmbH, Munich, Germany) with 63 Ag/AgCl electrodes placed at the standard positions of the international 10-20 system. Channels were referenced to the nose, with the ground placed at the frontal area around the AFz electrode. Impedances were kept below 10 kΩ. Electrooculogram (EOG) signals were also recorded. Signals were amplified and sampled at 1 kHz using two 32 channel amplifiers (Brain Amp by Brain Products, Munich, Germany).

The experiment was implemented in Python using the open-source BCI framework Pyff [11] with Pygame (http://pygame.org) and VisionEgg [12]. Data analysis and classification were performed with MATLAB (The MathWorks, Natick, MA, USA) using an in-house BCI toolbox (www.bbci.de/toolbox).

Participants were instructed to sit comfortably in a chair with a distance of 1 m between their eyes and the screen.

There were five conditions in total, with 12 distinct symbols for each. To choose a symbol, the participants needed to conduct a two-step selection procedure, resulting in (12*2 selections) 24 total selections.

We chose 12 distinct symbols such that, in the 24 target selections required to choose these 12 symbols, each of the six targets was to be chosen exactly four times. Then, for each condition, we randomly ordered these 12 letters/symbols and split them into two groups of six. The participant underwent ten total phases of the experiment, each phase requiring them to "spell" one of these groups of six letters/symbols for a specific condition; two groups of six for each of the five conditions give ten phases. We randomly ordered the phases with the constraint that the first five phases contained exactly one phase per condition, and the same for the second five phases. This random ordering process was performed separately for each participant. Each phase lasted about 3 minutes and 20 seconds and the participants were given breaks at their leisure between each one to ensure high focus during each phase.

We compared the temporal, spatial and overall classification accuracy of these conditions using regularized Linear Discriminant Analysis (LDA) with shrinkage of the covariance matrix [15]. For temporal accuracy, the data of a time window of 20 ms width with a step size of 10 ms of all the 63 channels were used as the feature. The whole time interval ([0 – 800 ms]) data of each EEG channel respectively were used as temporal feature to obtain the spatial classification. However, a realistic estimate of the spelling performance is obtained for classification with overall features (all channels and whole time intervals were used), cf. [15]. We calculated the grand averaged accuracies of different conditions along repetitions of stimuli. One repetition stands for one random presentation of all six visual and auditory stimuli.

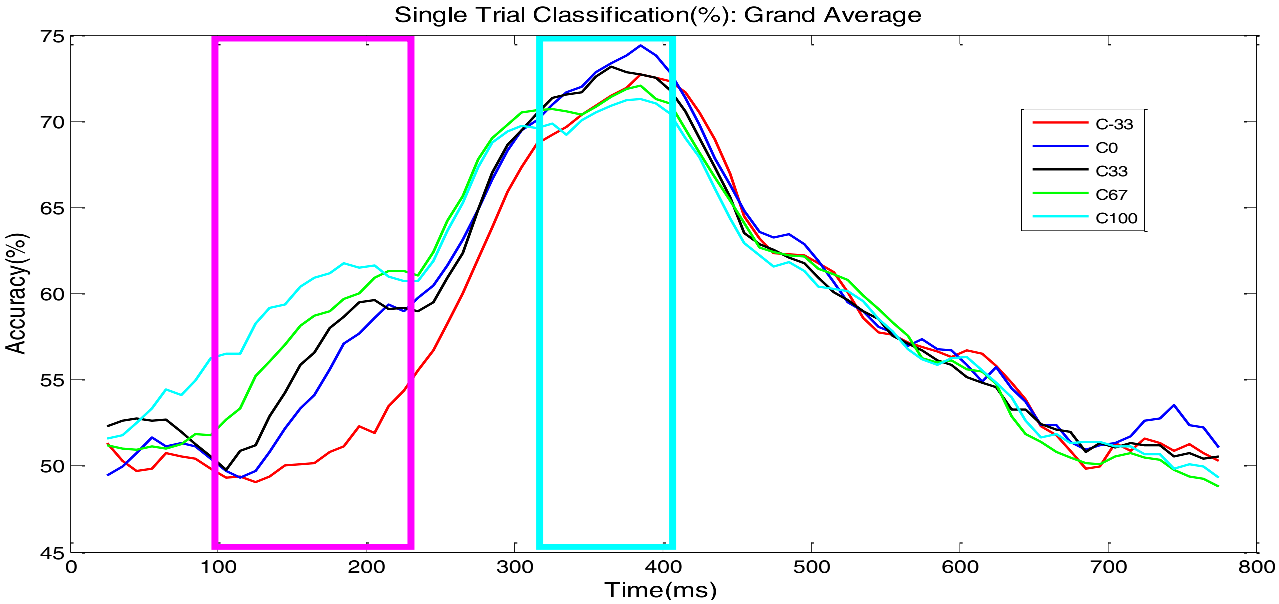

Figure 2. Temporal offline binary classification accuracies

## III. Results

### A. Temporal offline binary classification accuracies

Fig. 2 shows the temporal distribution of discriminative information. We observe that the accuracy varies greatly in the early temporal regions which represent the early negative ERP components from 100 to 200 ms. Classification accuracies of the early components increased as the visual-to-auditory delay increased. The highest single trial classification accuracy during 100 to 200 ms is obtained in condition 'C100' (over 60%), followed by conditions 'C67', 'C33' and 'C0'. Condition 'C-33' has the lowest accuracy at the chance level of 50%.

### B. Spatial distribution of the classification over channels

In Fig. 3, we displayed those accuracies per channel as scalp topographies as indication of the spatial distribution of discriminative information, cf [15]. Condition 'C33' has high accuracy in the central area (corresponding to the cognitive P3 component), while the parietal and occipital area did not show any superiority. In contrast, condition 'C-33' has higher accuracy in parietal and occipital areas (sensory visual ERP components) than other conditions.

### C. Overall classification accuracies over repetitions

Fig. 4 shows the grand averaged accuracies of different conditions along repetitions of stimuli. The red line (representing condition 'C-33') shows the best performance. The conditions with the highest visual-to-auditory stimuli delays ('C67' and 'C100') give the worst performance of all conditions.

## IV. Discussion

Campanella [1, 13] suggested that a cross-modal oddball design should be used in future studies to increase the sensitivity of the P300 amplitude differences between healthy participants and those with clinical symptoms. Hessler [14] also found that congruent audiovisual stimuli elicited an N2 response with a shorter latency and a P3 with smaller amplitude than auditory stimuli and showed that the whole is more than the sum of its parts in audiovisual processing. Senkowski [8] has discussed in his study that the precision of temporal synchrony can have an impact on early cross-modal interactions in human cortex. In our study, we spatially compared conditions with different visual-to-auditory onset delays, ranging from -33 ms to 100 ms, which are coded as conditions 'C-33','C0','C33','C67' and 'C100'.

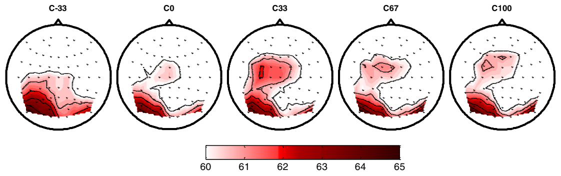

Figure 3. Spatial distribution of the classification over channels

The results of the performance are complex. From the temporal distribution of discriminative information, we find that the accuracy of the early time intervals during 100 to 200 ms is increased with prolonged visual-to-auditory delay. Though condition 'C100' has the best early component accuracy, and 'C-33' has the worst, the accuracies obtained for spatio-temporal features give the opposite picture.

One contribution to the high classification accuracy of 'C-33' could be the brain response in parietal/occipital areas (Fig 3). However, another explanation for these seemingly contradicting results of 'C-33'and 'C100' could be that the discriminative information in condition 'C-33' is contained in the combination of early and later components and different spatial regions. Thus scattered information is invisible when investigating the temporal and the spatial domain separately as in Figs 2 and 3, as it can only be exploited with spatio-temporal features as in Fig 4.

A possible reason for the better performance of 'C-33' could be the short real-time visual-to-auditory delay.

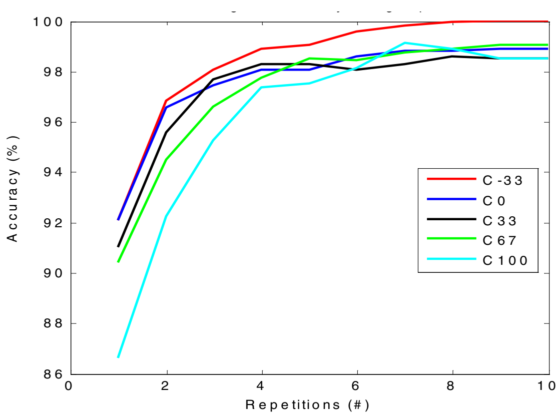

Figure 4. Grand-averaged accuracy over repetitions

Condition 'C100' has a long real-time delay, which means that at the onset of the visual stimuli, the auditory stimuli has almost finished. This could explain why the higher delays have less accuracy, since there is less overlap between the two stimulus modalities.

The results shown above indicate the need for further investigation of the complex interaction in multimodal stimulation paradigms and in particular conditions in which visual stimuli predate auditory stimuli.

## V. Conclusion

The brain responses to visual-auditory stimuli with different visual-to-auditory delays are complex. The tendencies of parietal and central parts are different for different visual-to-auditory delays. However, the contradictory results of single trial classification and the overall classification with different repetitions inspire us to investigate further studies to achieve better explanations and ultimately determine the ideal delay time for optimal BCI application.

## Acknowledgment

This work was supported by the China Scholarship Council and National Natural Science Foundation of China (Grant No. 81222021). Furthermore, we acknowledge financial support by the BMBF Grant No. 16SV5839. This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein.

## References

[1] S. Campanella, R. Bruyer, S. Froidbise, M. Rossignol, F. Joassin, C. Kornreich, P. Verbanck, "Is two better than one? A cross-modal oddball paradigm reveals greater sensitivity of the P300 to emotional face-voice associations," *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology.*, vol. 121 no. 11, pp. 1855–62, 2010.

[2] J.R. Wolpaw and E.W. Wolpaw, *Brain-Computer Interfaces: Principles and Practice*. Oxford University Press, 2012.

[3] G. Dornhege, *Toward Brain-Computer Interfacing*. MIT Press, 2007

[4] J.N. Mak and J.R. Wolpaw, "Clinical applications of brain-computer interface: current state and future prospects," *IEEE Rev. Biomed. Eng.,* vol. 2, pp.187-199, 2009.

[5] D.S. Klobassa, T.M. Vaughan, P. Brunner, N.E .Schwartz, J.R. Wolpaw, C. Neuper and E.W. Sellers, 2009 "Toward a high-throughput auditory P300-based brain-computer interface," *Clin Neurophysiol.* vol. 120, no. 7, pp. 1252-61, 2009.

[6] S. Boll, and S. Berti, (2009). "Distraction of task-relevant information processing by irrelevant changes in auditory, visual, and bimodal stimulus features: a behavioral and event-related potential study," *Psychophysiology*, vol. 46, no. 3, pp. 645–654, 2009.

[7] P. M. Jaekl, and L. R. Harris, " Auditory-visual temporal integration measured by shifts in perceived temporal locations." *Neuroscience Letters*, vol.417, pp. 219–224, 2007.

[8] D. Senkowski, D. Talsma, and M. Grigutsch, "Good times for multisensory integration: effects of the precision of temporal synchrony as revealed by gamma-band oscillations." *Neuropsychologia,* vol. 45, no.3, pp. 561-571, 2007.

[9] J. Höhne, K. Krenzlin, S. Dähne and M. Tangermann, "Natural stimuli improve auditory BCIs with respect to ergonomics and performance," *J.Neural Eng,* vol. 9, no. 4, 2012.

[10] M.S. Treder, N.M. Schmidt and B. Blankertz, "Gaze-independent brain–computer interfaces based on covert attention and feature attention," *J. Neural Eng,* vol. 8, no. 6, 2011.

[11] B. Venthur, S. Scholler, J. Williamson, and S. Dähne, "Pyff---A Pythonic Framework for Feedback Applications and Stimulus Presentation in Neuroscience," *Front Neurosci.* vol. 4, Dec 2010.

[12] Andrew D. Straw, "Vision Egg: An Open-Source Library for Realtime Visual Stimulus Generation." *Frontiers in Neuroinformatics.* vol. 2, no. 4, Nov 2008.

[13] S. Campanella, D. Delle-Vigne, C. Kornreich, and P. Verbanck, "Greater sensitivity of the P300 component to bimodal stimulation in an event-related potentials oddball task," *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, vol. 123, no. 5, pp. 937–46, 2012.

[14] D. Hessler, R. Jonkers, L. Stowe, and R. Bastiaanse, "The whole is more than the sum of its parts - Audiovisual processing of phonemes investigated with ERPs," *Brain and language*, vol. 124, no. 3, pp. 213–224, 2013.

[15] B. Blankertz, S. Lemm, M.S. Treder, S. Haufe, K.-R. Müller, Single-trial analysis and classification of ERP components - a tutorial. Neuroimage, 56:814-825, 2011