Speech Perception: Single Trial Analysis of the N1/P2 Complex of Unimodal and Audiovisual Evoked Responses

George Zouridakis, *Senior Member, IEEE,* Martijn Baart, Jeroen J. Stekelenburg, and Jean Vroomen

***Abstract*—Audiovisual speech integration is reflected in the electrophysiological N1/P2 complex. In this study, we analyzed recordings of electroencephalographic brain activity from 28 subjects who were presented with combinations of auditory, visual, and audiovisual stimuli, using single trial analysis based on an independent component analysis procedure. We found that, with respect to the N1/P2 complex, single trials across all subjects and all conditions could be separated into two groups, one with 'typical' responses having the same polarity as the average response and another one with 'aberrant' responses having opposite polarity. Furthermore, the number of aberrant responses in non-speech interpretation of auditory stimuli was significantly lower compared to speech, which may affect the N1/P2 complex in the ensemble average EP.**

I. Introduction

HUMAN speech is often audiovisual (AV) in nature as the auditory speech signal (A) is typically accompanied by visual lip-read information (V) in most natural settings. To determine AV interactions in speech processing, many studies use electroencephalographic (EEG) recordings of brain activity while listeners are presented with AV, A, and V speech stimuli [1,2-5], a procedure that allows researchers to interpret the resulting evoked potential (EP) differences between (AV–V) and A as an effect of AV integration [6,7,8]. In particular, such studies have shown that the auditory N1 (a negative peak approximately 100 msec after stimulus onset) and the P2 (a positive peak approximately 200 msec poststimulus) EP components are modulated by lip-read speech [1-5].

The typical procedure to analyze EPs is to ensemble average all responses obtained under each experimental condition, and then compare the polarity, amplitude, and latency of the resulting EP components across conditions. Averaging improves the signal-to-noise ratio in the recorded signals, which are invariably affected by extraneous activity, such as physiological artifacts and external noise. However, averaging does not provide information about the dynamics of the brain processes underlying the surface recordings. Single-trial analysis, on the other hand, can provide information on the temporal evolution of the neurophysiological processes associated with the particular EP components under investigation. We have recently proposed a methodology for single-trial EP analysis [17] that focuses only on one particular EP component at a time and makes it visible in each single trial. Our method is based on independent component analysis (ICA) and the idea that activity resulting from an experimental stimulus is independent from neurophysiological artifacts and background brain activity [14,15,16]. The method allows studying the dynamic evolution of the underlying cortical generators that give rise to a specific EP component.

In previous studies, the single-trial ICA approach revealed that a sub-set of single-trial EP components have an opposite polarity than the one observed in the ensemble average [9,12,13,17]. Since these atypical or 'aberrant' responses can distort the effective morphology of the measured average EP components as well as the location of the underlying intracranial sources [9], they should be taken into account in the analysis of any brain process under investigation.

In this study, we employed unimodal (A or V) and bimodal (AV) stimulation, whereby listeners were presented with artificial sine-wave speech (SWS) [10] that was only perceived as speech by half of the participants and we investigated whether the N1/P2 complex is modulated by the speech- or non-speech interpretation of the stimuli.

II. Methods

A. Subjects and Procedure

Participants (28 undergraduate students, 8 males, 20 females, age between 18 and 26 years) were seated in an electrically shielded, dimly lit, and sound attenuated booth at ~70 cm from a 17-inch CRT monitor. SWS sounds, specifically the pseudo-words /tabi/ and /tagi/, were delivered through a computer speaker placed directly below the monitor. For visual stimuli, the size of the videos viewed subtended 14° horizontal and 12° vertical visual angle.

Participants were evenly split (N=14) into a 'speech mode' (SM) and a 'non-speech mode' (NSM) group and trained accordingly. SM participants learned to perceive the SWS as speech through alternate presentations of SWS and their natural speech counterparts (twelve presentations of each stimulus). Listeners in NSM only heard the SWS sounds (also 12 times for each sound) while under the impression they were hearing two computer sounds. After training, none

Manuscript received on July 20, 2013.

G.Z. is Visiting Professor at the Basque Center on Cognition, Brain and Language, San Sebastián, Spain. His permanent appointment is with the Departments of Engineering Technology, Computer Science, and Electrical and Computer Engineering, University of Houston, Houston, TX 77204, USA(+1-713-743-8656; fax: +1-713-743-0172; e-mail: zouridakis@uh.edu).

M.B. is with the Basque Center on Cognition, Brain and Language, San Sebastián, Spain (e-mail: m.baart@bcbl.eu).

J.J.S. and J.V. are with the Department of Cognitive Neuropsychology at Tilburg University, Tilburg, The Netherlands (e-mail:j.j.stekelenburg@uvt.nl and j.vroomen@uvt.nl).

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

of the SM participants reported to have heard the sounds as speech. Next, EPs were recorded during six 10-minute blocks with short breaks in between. Across blocks, a total of 576 experimental trials were delivered, of which 288 were unimodal (144 A- and 144 V-trials, 72 /tabi/ and 72 /tagi/ trials in each modality) whereas 288 were AV. Furthermore, 144 AV trials were congruent (AVC with 72 AV/tabi/ and 72 AV/tagi/) and 144 were incongruent (AVI with 72 A/tabi/V/tagi/ and 72 A/tagi/V/tabi/). During the experiment, all participants were engaged in an unrelated visual task.

B. *EEG recordings*

EEG activity was recorded at a sampling rate of 512 Hz from 64 locations corresponding to the extended International 10-20 system, using active electrodes (BioSemi, Amsterdam, The Netherlands) mounted in an elastic cap. The EEG was referenced on-line through two additional electrodes; the active Common Mode Sense electrode (CMS) and ground (Driven Right Leg passive electrode; DRL). Electrooculographic (EOG) activity was recorded using 4 additional electrodes (2 on the orbital ridge above and below the right eye and 2 on the lateral junctions of both eyes) referenced to the left and right mastoids. The EEG was referenced offline to an average of these mastoids and bandpass filtered (Butterworth zero-phase filter, 0.5 – 30 Hz, 24 dB/octave). EPs were time-locked to auditory onset and the raw data were segmented into epochs of 1000 msec, including a 200-msec pre-stimulus baseline. After EOG correction [11], epochs with an amplitude > 150 μV at any EEG channel were rejected.

C. *Iterative ICA*

Independent component analysis [3] is a method for solving the blind source separation problem [7], which tries to recover N independent source signals, $s = \{s_1,...,s_N\}$, from N observations, $x = \{x_1,...,x_N\}$, that represent linear mixtures of the independent source signals. The key assumption used to separate sources from mixtures is that the sources are statistically independent, while the mixtures are not. Mathematically, the problem is described as $x = As$, where A is an unknown mixing matrix, and the task is to recover a version, u, of the original sources, similar to s, by estimating a matrix, W, which inverts the mixing process, i.e., $u = Wx$. The estimates u are called independent components (ICs). The extended infomax algorithm is currently the most efficient technique to solve this problem and relies on information theory and a neural network approach [1,4,7,8].

Our technique, termed iterative ICA (iICA), is an iterative implementation of this algorithm and is applied to a set of recordings consisting of L single trials obtained from N recording channels. Before processing, all single trials are lowpass filtered at 35 Hz. The procedure consists in the following steps:

1. Compute an average EP from all single trials.
2. Compute the ICA transform of all single trials, grouped in blocks of 10.
3. Compute the absolute correlation value between the current average EP and the ICs in all blocks, within a predefined window W_r.

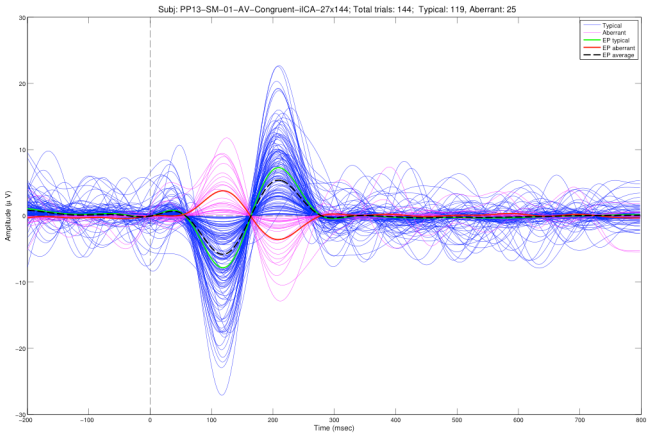

Fig. 1. Example of 'typical' and 'aberrant' single trial responses identified by the iICA algorithm, along with the resulting partial EPs. The classical ensemble average EP is also shown for comparison.

4. Set to zero those ICs with correlation less than a predefined threshold r_{th}.
5. Compute the inverse ICA transform of the updated ICs back to the time domain, separately in each block.
6. Shuffle the updated single trials around the entire set.
7. Repeat steps 1 to 6 until a convergence criterion is met.

The procedure can then be applied to the rest of the channels until all of them have been processed. However, in the present study we analyzed data only from the Cz channel as the N1/P2 complex in the raw data was maximized at Cz. The parameter values used were $W_r = 50 – 250$ msec poststimulus, which was consistent with the occurrence of the N1/P2, and $r_{th} = 0.15$. For convergence, we accepted an absolute difference between successive estimates of the template $|EP_k – EP_{k+1}| \leq 0.005$. Shuffling of the trials guarantees that each block will include different trials in the next iteration, and thus the resulting ICA system of equations will not be underdetermined.

D. *Analysis Procedure*

EPs were computed for each modality (A, V, AVC, and AVI), separately for the SM and NSM subgroups. The N1/P2 complex was identified automatically in a window between 50–250 msec poststimulus. In all datasets, the iICA procedure was able to identify two groups of single trials in which the N1/P2 had opposite polarity. When averaged together, the iICA-processed single trials of the same group resulted in two distinct partial EPs. We called 'typical' those responses in which the N1/P2-complex had the same polarity as in the ensemble average EP and 'aberrant' the ones in which N1/P2 had opposite polarity.

We determined the proportion of aberrant responses for each participant in all conditions and submitted the data to a 2 (Group; SM vs NSM) * 4 (Modality; A, V, AVC and AVI) ANOVA.

III. Results

A characteristic example of typical and aberrant responses, along with the corresponding partial EPs is shown in Fig. 1. The classical ensemble average EP, labeled 'EP average,' is also plotted in the same figure for comparison. As it can be seen, both the N1 and the P2 components have opposite polarities, and the main effect of the aberrant responses on the ensemble average EP is to decrease the amplitude of both the N1 and P2 waves.

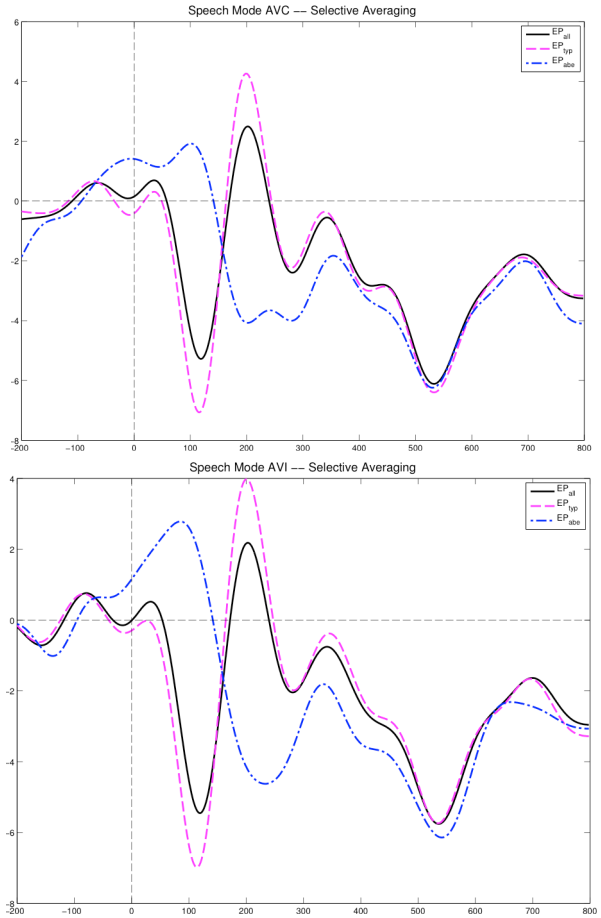

Fig. 2. Examples of selective averaging of congruent (top) and incongruent (bottom) audiovisual EPs in which unprocessed single trials are grouped and averaged together based on the membership label of 'typical' or 'aberrant' assigned by the iICA algorithm. The classical ensemble average EP is also shown for comparison.

Figure 2 shows two examples of selective averaging for congruent (AVC, Fig. 2, top) and incongruent (AVI, Fig. 2, bottom) EPs. In these plots, the original unprocessed single trial responses were first grouped and then averaged together based on the label ('typical' or 'aberrant') of the corresponding processed trials assigned by the iICA algorithm. Similar to what is seen in the iICA-processed single trials, both the N1 and P2 peaks have opposite polarity in the two partial EPs, whereas the polarity of later components is the same in all EPs. That is, components that are out-of-phase during the time window of the N1/P2 complex, approximately between 50 and 250 msec poststimulus, become in-phase at the end of this stage of processing. Similar results were found for the non-speech mode EPs.

Table 1 summarizes the proportion of single trials having aberrant N1/P2 polarity across the SM and NSM groups for A, V, AVC, and AVI stimulus modalities. The overall proportion of aberrant responses was larger for listeners in SM (.28) than in NSM (.22), $F = 16.37$, $p = .027$. Furthermore, there was a main effect of Modality as the proportion of aberrant responses was significantly higher for V than any other condition (p's < .001), whereas we observed no statistical difference between the other levels of Modality.

Table 1. Proportion of aberrant responses for all conditions in SM, NSM, and averaged across participants, with standard deviations in parentheses. Asterisks indicate significant differences where $p < .001$.

Modality	SM	NSM	Average	
A	.24 (.21)	.16 (.09)	.20 (.16)	*
V	.38 (.04)	.36 (.05)	.37 (.04)	*
AVC	.25 (.18)	.17 (.10)	.22 (.15)	*
AVI	.24 (.18)	.17 (.10)	.20 (.15)	

IV. Discussion & Conclusion

The results presented in this paper are in agreement with previous findings of P2 component amplitude differences during processing of audiovisual speech and non-speech, specifically that the P2 amplitude may be larger for non-speech (i.e., NSM in current study) than for speech (SM in current study) material [7]. Selective averaging revealed that the number of aberrant responses in non-speech interpretation of artificial stimuli is lower compared to speech, thus affecting the amplitude of the P2 component in the ensemble average EP. Our efforts are currently focused on localizing the source of the aberrant responses and defining the relative timing of the two partial EP peaks across the experiment.

Taken as a whole, these results suggest that single trial analysis can shed more light on the generation of the N1/P2 complex and its dynamic evolution.

Acknowledgment

This work was supported in part by an award to GZ from Ikerbasque, the Basque Foundation for Science in Spain.

REFERENCES

[1] J. Besle, A. Fort, C. Delpuech *et al.*, "Bimodal speech: early suppressive visual effects in human auditory cortex," *European Journal of Neuroscience,* vol. 20, no. 8, pp. 2225-34, Oct, 2004.

[2] J. J. Stekelenburg, and J. Vroomen, "Neural correlates of multisensory integration of ecologically valid audiovisual events," *Journal of Cognitive Neuroscience,* vol. 19, no. 12, pp. 1964-73, Dec, 2007.

[3] V. Klucharev, R. Möttönen, and M. Sams, "Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception," *Cognitive Brain Research,* vol. 18, no. 1, pp. 65-75, Dec, 2003.

[4] V. van Wassenhove, K. W. Grant, and D. Poeppel, "Visual speech speeds up the neural processing of auditory speech," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 102, no. 4, pp. 1181-6, Jan 25, 2005.

[5] J. J. Stekelenburg, and J. Vroomen, "Electrophysiological correlates of predictive coding of auditory location in the perception of natural audiovisual events.," *Frontiers in Integrative Neuroscience,* vol. 6, 2012.

[6] A. Fort, C. Delpuech, J. Pernier *et al.*, "Early auditory–visual interactions in human cortex during nonredundant target identification.," *Brain Research, Cognitive Brain Research,* vol. 14, pp. 20–30, 2002.

[7] M. H. Giard, and F. Peronnet, "Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study," *Journal of Cognitive Neuroscience,* vol. 11, no. 5, pp. 473-90, Sep, 1999.

[8] S. Molholm, W. Ritter, M. M. Murray *et al.*, "Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study," *Brain Research, Cognitive Brain Research,* vol. 14, no. 1, pp. 115-28, Jun, 2002.

[9] D. Iyer, J. Díaz, and G. Zouridakis, "Consistency of the auditory evoked response: the presence of aberrant responses and their effect on N100 localization," *J Neurosci Methods,* vol. 208, no. 2, pp. 173-80, Jul, 2012.

[10] R. E. Remez, P. E. Rubin, D. B. Pisoni *et al.*, "Speech perception without traditional speech cues," *Science,* vol. 212, pp. 947-9, May 22, 1981.

[11] G. Gratton, M. G. Coles, and E. Donchin, "A new method for off-line removal of ocular artifact," *Electroencephalography and Clinical Neurophysiology,* vol. 55, no. 4, pp. 468-84, 1983, 1983.

[12] D. Iyer, N.N. Boutros, G. Zouridakis "Single-Trial Analysis of Auditory Evoked Potentials Improves Separation of Normal and Schizophrenia Subjects," Clinical Neurophysiology, 2012 Sep; 123(9): 1810-20

[13] D. Iyer, and G. Zouridakis "Single-trial evoked potential estimation: comparison between independent component analysis and wavelet denoising," Clinical Neurophysiology, 118(3): 495-504, 2007.

[14] S. Makeig, M. Westerfield, T-P. Jung, J. Covington, J. Townsend, T. J. Sejnowski, E. Courchesne, "Independent components of the late positive response complex in a visual spatial attention task," J Neurosc,i 19: 2665–2680, 1999.

[15] S. Makeig, M. Westerfield, T-P. Jung, S. Enghoff, J. Townsend, E. Courchesne, T. J. Sejnowski, "Dynamic brain sources of visual evoked responses", Science, 295: 690-694, 2002.

[16] R. N. Vigario, "Extraction of ocular artifacts from EEG using independent component analysis", Electroenceph Clin Neurophysiol, 103: 395- 404, 1997.

[17] G. Zouridakis, D. Iyer, J. Diaz, U. Patidar, Estimation of individual evoked potential components using iterative independent component analysis, Physics in Medicine and Biology, 52(17): 5353-5368, 2007.