ELECTROPHYSIOLOGY OF UNIMODAL AND AUDIOVISUAL SPEECH PERCEPTION

Lynne E. Bernstein [1], *Curtis W. Ponton* [2] , *Edward T. Auer, Jr.* [1]

[1] House Ear Institute, 2100 W. Third St., Los Angeles, CA 90057. [2] Neuroscan Labs, 5700 Cromo Dr. – STE 100, El Paso, TX 79912.

ABSTRACT

Based on behavioral evidence, audiovisual speech perception is generally thought to proceed linearly from initial unimodal perceptual processing to integration of the unimodally processed information. We investigated unimodal versus audiovisual speech processing using electrical event-related potentials (ERPs) obtained from twelve adults. Nonsense syllable stimuli were presented in an oddball paradigm to evoke the mismatch negativity (MMN). Conditions were (1) audiovisual incongruent stimuli (visual /ga/ + auditory /ba/) versus congruent audiovisual stimuli (visual /ba/ + auditory /ba/), (2) visual-only stimuli from the audiovisual condition (/ga/ vs. /ba/), and (3) auditory-only stimuli (/ba/ vs. /da/). A visual-alone MMN was obtained on occipital and temporo-parietal electrodes, and the classical auditory MMN was obtained at the vertex electrode, Cz. Under audiovisual conditions, the negativity recorded at the occipital electrode locations was reduced in amplitude and latency compared to that recorded in the visual-only condition. Also, under the audiovisual condition, the vertex electrode showed a smaller negativity with increased latency relative to the auditory MMN. The neurophysiological evidence did not support a simple bottom-up linear flow from unimodal processing to audiovisual integration.

1. INTRODUCTION

Speech perception is a process that transforms speech signals into the neural representations that are then projected onto word-form representations in the mental lexicon. *Phonetic perception* is more narrowly defined as the perceptual processing of the linguistically relevant attributes of the physical (measurable) speech signals. How acoustic and optical phonetic speech signals are processed under unimodal conditions and integrated under audiovisual conditions are fundamental questions for behavioral and brain sciences. The McGurk effect [1] has been used as a tool in investigating this question behaviorally. An example of the McGurk effect is when a visible spoken token [ga] is presented synchronously with an audible [ba], frequently resulting in the reported percept /da/.

Various experimental results have led to the McGurk effect being attributed to, and viewed as evidence for, early bottom-up perceptual integration of phonetic information. For example, selectively attending to one modality only does not abolish the effect [2], suggesting that cognitive top-down control is not possible. Gender incongruency and knowledge about phonemic incongruency between auditory versus visual stimuli does not abolish it [3, 4], suggesting a high degree of bottom-up automaticity. Auditory phonetic distinctions, for example voicing, are affected by visual syllables [5], and phonetic goodness judgments are affected by visual syllables [6], suggesting that integration is subphonemic. McGurk percepts do not adapt auditory stimuli [7], suggesting that audiovisual integration strictly follows auditory phonetic processing. These observations are consistent with the theory that there is a bottom-up flow of unimodal processing that precedes audiovisual integration. There are also results inconsistent with the strictly bottom-up flow of unimodal information followed by integration. For example, asynchrony of audiovisual stimuli of approximately 180 ms does not abolish McGurk effects [cf., 8, 9], suggesting that perceptual information can be maintained in memory and then integrated. Reductions in effect strength have also been shown to occur due to training [2] and to talker familiarity [10], both of which may be related to post-perceptual processes.

2. THE CURRENT STUDY

We investigated the processing of unimodal auditory and visual, and audiovisual speech stimuli using recordings of electrical event-related potentials (ERPs) obtained in an oddball, mismatch negativity (MMN) paradigm. ERPs afford neurophysiological measures of brain activity with high temporal resolution (< 1 ms) and with moderately good spatial resolution (< 10 mm). These measures of thalamic and cortical brain activity are presumed to reflect mostly excitatory post-synaptic potentials arising from large populations of pyramidal cells oriented in a common direction [12-14]. ERPs are often classified as exogenous (reflecting physical stimulus characteristics, such as intensity) or endogenous (reflecting the state of the subject or the cognitive demands of the task; [15]).

The cortical auditory ERP N1 is an exogenous, obligatory potential that occurs approximately 100 ms after stimulus onset. It is preceded by the P1, at approximately 40-50 ms, and followed by the P2, at approximately 200 ms after stimulus onset. The auditory N1 varies with physical stimulus characteristics and rate [15]. In an oddball presentation paradigm (e.g., the standard /ba/ repeated for 83% of trials and /da/ 17% pseudorandomly interspersed as the deviant), N1 is elicited by the standard and the deviant [e.g., 16], but the deviant can generate an additional negativity described as the mismatch negativity (MMN). The MMN is considered to be an automatic preattentive difference detection response correlated with behavioral discrimination (e.g., [16]). The MMN is isolated from the obligatory potentials by subtraction of the ERP to a stimulus playing the role of deviant from the ERP to the *same* stimulus playing the role of the standard. This controls for stimulus effects and gives the contrast, or change detection, effect only. In this study, MMN was defined as an added negative component evoked by an infrequently occurring deviant stimulus in a sequence of repeating standard stimuli, regardless of perceptual modality. In addition, if the magnitude of the difference between the ERP to the stimulus in its roles as standard versus deviant is sufficiently large, then a positive-going potential known as the P300 may be present and partially superimposed on the MMN. This can be seen in the response to the deviant in the top panel of Fig. 1 for the auditory stimulus /ba/.

Previously, Sams et al. obtained event-related magnetoencephalographic recordings in response to congruent and incongruent audiovisual speech stimuli, which they interpreted as evidence that visual phonetic information modulates responses to auditory stimuli by the auditory cortex [11]. This interpretation is consistent with non-linearity in bottom-up processing. Sams et al. did not obtain a visual-only response. We sought to replicate and extend their previous research.

3. METHODS

Subjects. Twelve adults participated for pay in the experiment. All had normal hearing, normal or corrected-to-normal vision, English as a first language, and susceptibility to the McGurk effects, as verified with behavioral testing.

Stimuli. The stimuli were spoken tokens of /ba/, /da/, and /ga/. They were digitally recorded so that the visual start of each token continued seamlessly from the end of the other tokens. The /ba/ and /da/ tokens were used in the auditory conditions. The /da/ stimulus was used because it is the typical percept obtained with the McGurk stimulus that comprises visual /ga/ and auditory /ba/. The visual stimuli were /ba/ and /ga/. The audiovisual stimuli were congruent audiovisual "/ba-ba/", and incongruent auditory /ba/ combined with visual /ga/ ("/ba-ga/"). Thus /ba/ was constant during audiovisual trials.

The acoustic stimuli were 500 ms in duration. The video tokens were 20 frames (666 ms). The visual stimuli diverged from each other approximately 2-3 frames from their onset. The acoustic tokens were aligned at their onset with either the original acoustic /ba/ burst (the congruent stimuli) or the original acoustic /ga/ burst (the incongruent stimuli), approximately half-way through Frame 7. The visual tokens therefore showed speech movement prior to the acoustic onset, as is normally the case.

ERP recording. A Neuroscan®, evoked potential recording system was used with electrode positions: C1, C2, C3, C4, C5, C6, Cz, F3, F4, F5, F6, F7, F8, Fz, T3, T4, T5, T6, P3, P4, P5, P6, PZ, O1, O2, Oz, M1, M2, Fp1, and Fp2. Ocular movements were monitored on two differential recording channels. A signal controlling stimulus presentation triggered sampling of the EEG. The EEG was anti-aliased, sampled at 1.0 kHz, and bandpass-filtered from either DC or 0.1 Hz to 300 Hz. Individual EEG epochs were recorded and stored to computer hard-disk for off-line processing. Off-line, single sweeps were baseline-corrected on the pre-stimulus interval. Using an automatic rejection algorithm, sweeps containing activity in any channel that exceeded ±200 µV were excluded from subsequent averaging. A regression-based correction for eye movements was applied to the accepted data. The 0-ms point of the ERP recordings was set at the onset of the acoustic stimuli. Sampling preceded the 0-ms point by 100 ms and followed for 600 ms.

MMN procedure. Each stimulus was tested as both a standard and a deviant. Standards were presented in 87% of trials pseudorandomly ordered with 13% of deviant trials. For example, auditory /ba/ occurred as a standard versus auditory /da/ as a deviant, and vice versa in another run. Thus there were six runs, two for each condition. 2200 trials were presented per subject per condition. During the auditory-only stimulus presentations, subjects watched a video of a movie (entertainment—not the visual syllable stimuli) with the sound off. Visual stimuli were viewed at a distance of 1.9 m from the screen. Testing took approximately 4.5 hours per subject.

Analyses. Grand mean waveforms were computed separately for the standard stimulus and the deviant stimulus in each condition, following initial response processing. The MMN was obtained by subtracting the standard response from the deviant response within stimulus and condition. An integrated MMNi [17] was derived from MMN difference waveforms and is shown in all of the figures below.

4. RESULTS

Fig. 1 shows the auditory MMNi at the central electrode Cz, which was in the range of typical auditory latencies (191 msec for /da/ and 220 msec for /ba/). Fig. 1 also shows the corresponding Cz responses for the other conditions. An MMNi was present at Cz for the visual /ga/ condition. Only a small MMNi was observed at Cz for the visual /ba/ and for audiovisual stimuli. Visual and audiovisual conditions resulted in longer MMNi latencies than did the auditory condition.

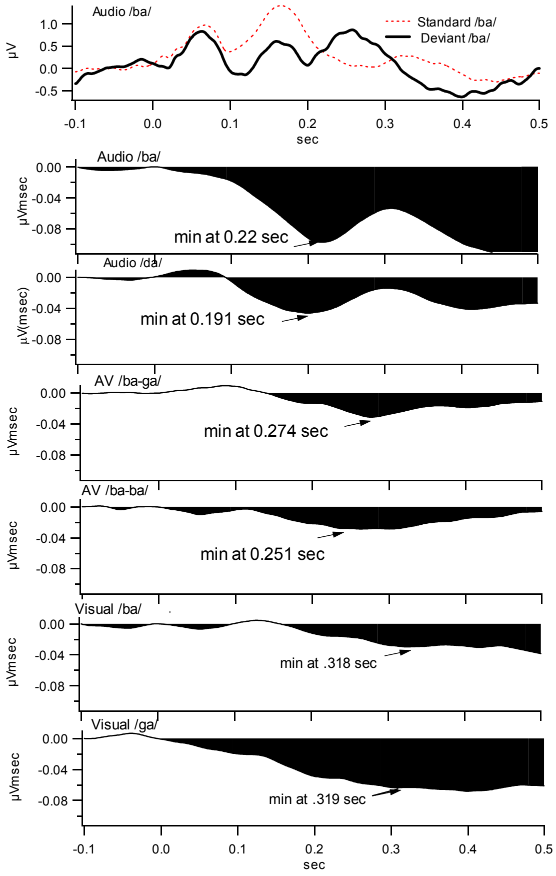
Fig. 2 shows ERPs evoked by the standard and deviant visual stimuli. Electrode sites in Fig. 2 were selected for display, because they best demonstrated the visual MMN. Beneath each waveform plot is shown the corresponding MMNi waveforms. The visual /ba/ condition generated a robust MMN at the occipital electrodes. This negativity was strongly right-lateralized at the temporal and parietal electrodes. The visual /ga/ condition generated an added negativity largest at occipital electrodes, without an easily definable peak latency, except perhaps at electrode O1.

Fig. 3 shows the results for the audiovisual congruent /ba-ba/ and the incongruent /ba-ga/. The MMNi activity for congruent /ba-ba/ was smaller in magnitude (-.02 vs. -.05 µV) and earlier in latency (114 vs. 157 msec) relative to the responses measured at occipital electrodes for visual /ba/ in Fig. 2. The /ba-ga/ stimulus resulted in an MMN-like response on the Oz and O2 electrodes, with peak negativity at 212 msec. The responses at the electrodes shown in Fig. 3 demonstrate that when a constant auditory /ba/ was presented, the responses at the occipital and temporo-parietal electrodes were altered relative to the responses in the visual-alone conditions.

5. SUMMARY AND CONCLUSIONS

These neurophysiological results revealed a more complex pattern of cortical activation than could be described as merely linearly ordered unimodal processing followed by audiovisual integration. The audiovisual /ba-ba/ MMNi on occipital electrodes was earlier and diminished in amplitude under conditions of the constant auditory /ba/, in comparison with the visual-only /ba/ response. The audiovisual MMNi response at vertex electrode Cz was later and smaller than the auditory-only MMNi. Analyses that space here did not permit discussing showed additional distinctly different response patterns across congruent versus incongruent audiovisual stimuli.

Figure 1: Waveforms for classical MMN to auditory /ba/ on electrode Cz (top panel). MMNi at Cz for each of the stimuli, bottom six
panels.



6. ACKNOWLEDGMENTS

The authors gratefully acknowledge the contributions made by Betty Kwong in collecting the electrophysiology data, and Sven Mattys, Ph.D., in assisting in the development of the stimuli and screening the subjects. Research supported by the NSF (9996088).

7. REFERENCES

1. McGurk, H. and MacDonald, J. “A visible production of a consonant /g/, presented simultaneously with a heard /b/, resulted in the percept /d/,” *Nature* 264: 746-748, 1976.

2. Massaro, D. W. *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Erlbaum Hillsdale, NJ, 1987.

3. Green, K. P. et al. "Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect," *Percept. Psychophys*. 50: 524-536, 1991.

4. Summerfield, Q. and McGrath, M. "Detection and resolution of audio-visual incompatibility in the perception of vowels," *Quart. J. Exp. Psych.* 36A: 51-74, 1984.

5. Green, K. P. and Kuhl, P. K. "The role of visual information in the processing of place and manner features in speech perception," *Percept. Psychophys.* 45: 34-42, 1989.

6. Brancazio, L., Miller, J. L., and Pare, M. A. "Visual influences on internal structure of phonetic categories," *J. Acoust. Soc. Am*. 108: 2481, 2000.

7. Saldana, H. M., and Rosenblum, L. D. "Selective adaptation in speech perception using a compelling audiovisual adaptor," *J. Acoust. Soc. Am.* 95: 3658-3661, 1994.

8. Massaro, D. W., Cohen, M. M., and Smeele, P. M. T. "Perception of asynchronous and conflicting visual and auditory speech," *J. Acoust. Soc. Am.* 100: 1777-1786, 1996.

9. Munhall, K. G., Gribble, P., Sacco, L., and Ward, M. "Temporal constraints on the McGurk effect," *Percept. & Psychophys*. 58: 351-362, 1996.

10. Walker S., Bruce, V., and O'Malley, C. "Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect," *Percept. Psychophys.* 57: 1124-1133, 1995.

11. Sams, M., Aulanko, R. Hamalainen M. et al. Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex. *Neurosci. Let*. 127: 141-145, 1991.

12. Creutzfeldt, O. D., Watanabe, S., and Lux, H. D. Relations between EEG phenomena and potentials of single cortical cells. I. Evoked responses after thalamic and epicortical stimulation. *EEG Clin. Neurophys*. 20**:** 1-18, 1966.

13. Miztdorf, U. "The physiological causes of VEP: Current source density analysis of electrically and visually evoked potentials," in *Evoked Potentials* (Eds, R. Q. Cracco, I. Bodis-Wollner), Alan R. Liss Inc, NY, 1986.

14. Vaughan, H., and Arezzo, J. The neural basis of event-related potentials," in *Human event-related potentials* (Ed. T. W. Picton), Elsevier, Amsterdam, 1988.

15. Näätänen R. and Picton, T. W. "The N1 wave of the human electric and magnetic response to sound: review and analysis of the component structure," *Psychophys.* 24: 375-425, 1987.

16. Näätänen R. "The mismatch negativity: A powerful tool for cognitive neuroscience," *Ear Hear.* 16: 6-18, 1995.

17. Ponton, C. W., Don, M., Eggermont, J. J., and Kwong, B. "Integrated mismatch (MMNi): A noise free representation that allows distribution free single-point statistical tests," *EEG Clin. Neurophys.* 104: 143-150, 1997.

Figure 2. Visual responses to /ba/ (left column) and /ga/ (right column). Top two rows show occipital electrode sites. Bottom two rows show temporo-parietal sites.

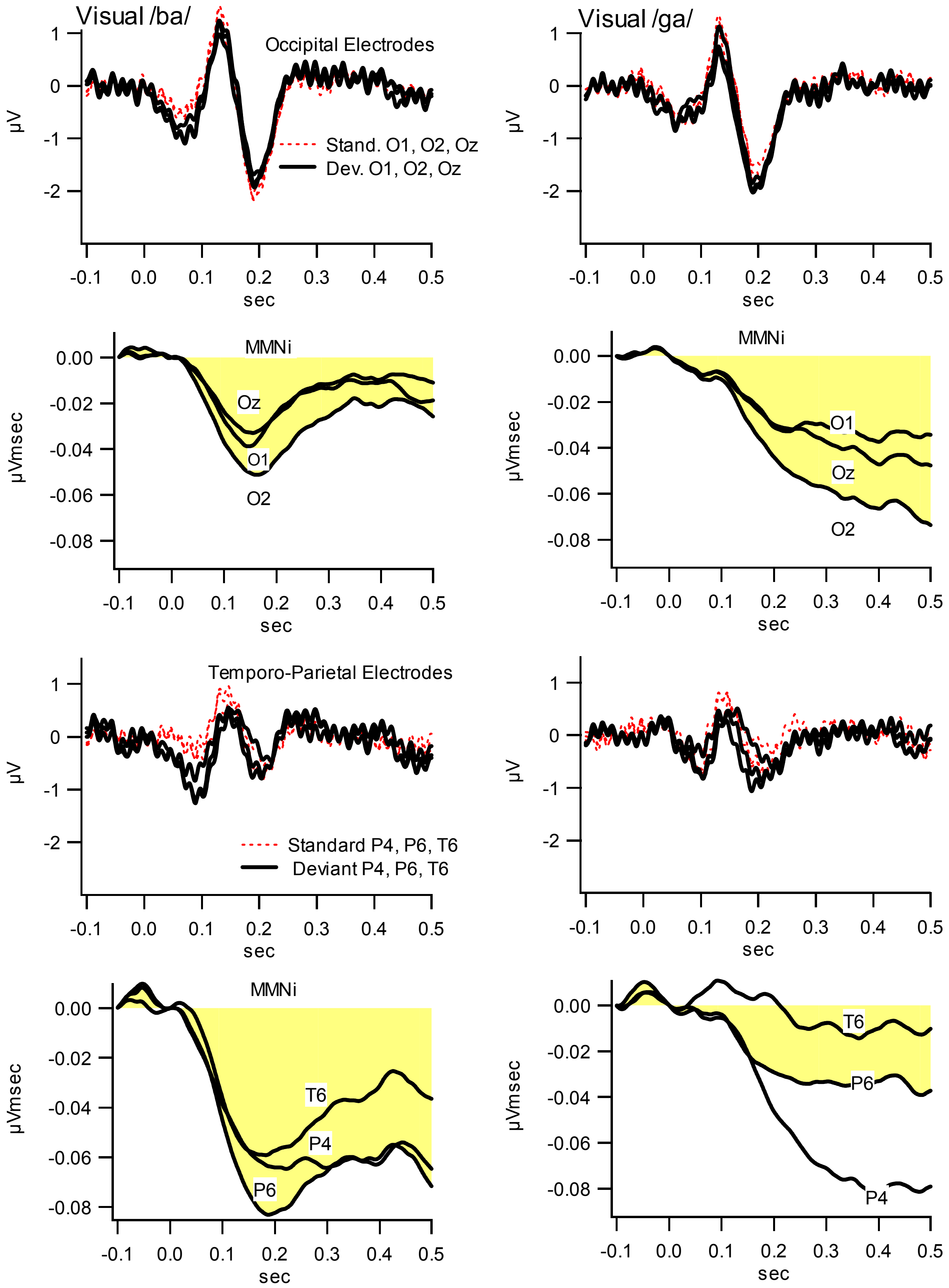
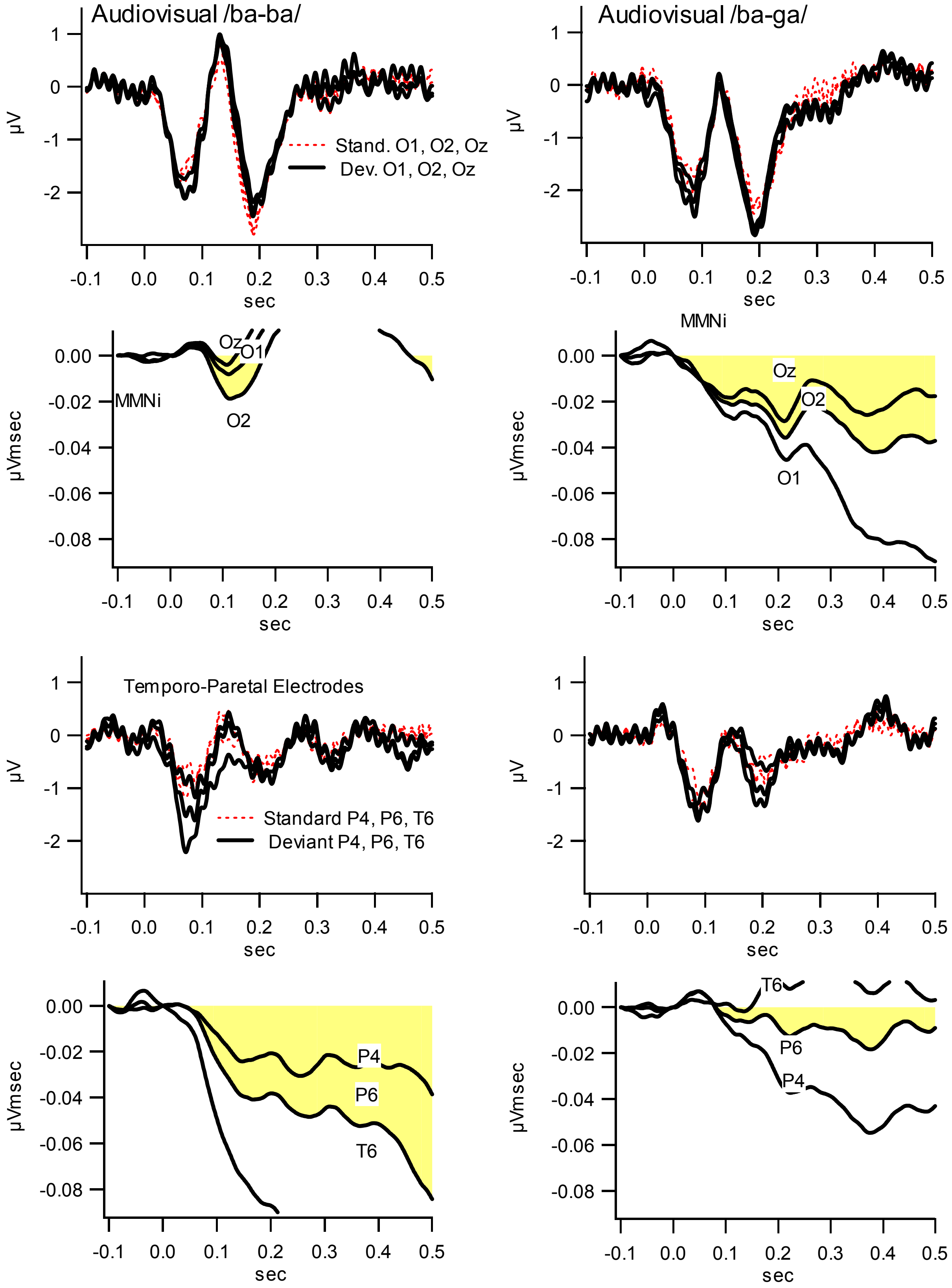


Figure 3. Audiovisual responses to /ba-ba/ (left column) and /ba-ga/ (right column). Top two rows show occipital electrode sites. Bottom two rows show temporo-parietal sites.


Audiovisual /ba-ba/
µV
Stand. O1, O2, Oz
Dev. O1, O2, Oz
sec
Audiovisual /ba-ga/
µV
sec
µVmsec
MMNi
Oz
O1
O2
sec
MMNi
µVmsec
Oz
O2
O1
sec
µV
Tem poro-Paretal Electrodes
Standard P4, P6, T6
Deviant P4, P6, T6
sec
µV
sec
µVmsec
P4
P6
T6
sec
µVmsec
T6
P6
P4
sec