ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Audiovisual Integration of Unfamiliar Words Reflects Multisensory Interaction in Phonetic Perception

Wichian Sittiprapaporn*

Faculty of Medicine, Mahasarakham University, Maha Sarakham, Thailand

ABSTRACT

Learning the correspondences between letters and speech sounds of a language is a crucial step in reading acquisition. The perception of speech and the inherently linked lip movements (audiovisual speech) emerged simultaneously during evolution, shaping the brain for integrating this audiovisual information. This study was conducted electroencephalography (EEG) experiments to characterize the influence of visual orthography on the most robust auditory event-related potentials (ERPs) and focused the analysis on systematic variation of the auditory ERPs as a function of visual orthography information. The subjects received auditory, visual, and audiovisual letters and were required to identify them, regardless of stimulus modality. Audiovisual letters included matching letters, in which the auditory and visual stimulus corresponded to each other based on previous experience, and nonmatching (randomly paired) letters. Meaningless auditory, visual, and audiovisual control stimuli were presented. The results showed that both non-phonetic and phonetic audiovisual interactions were found in the ERPs similar to the AV stimuli. The differences in the sum of the ERPs to the unimodal A and V stimuli and in ERPs to AV stimuli indicated interactions presumably based on temporal of the A and V components of the AV stimuli. The differences in the ERPs to the meaningful and meaningless of AV stimuli probably reflect multisensory interactions in phonetic processing.

Keywords: Event-related potentials; Tonal language; Lexical word; Audiovisual integration; Perception

***Corresponding author**

INTRODUCTION

Reading is essential to social and economic success in the present technological society [1]. In contrast to spoken language, which is a product of biological evolution, reading and writing are cultural inventions from the last few thousand years and are only relevant for most people since a few hundred years [2]. An intriguing question is, therefore, how it is possible that most people acquire literacy skills with such remarkable ease even though a naturally evolved brain mechanism for reading is unlikely to exist. An interesting hypothesis is that evolutionarily adapted brain mechanisms for spoken language provides a neural foundation for reading ability, which is illustrated by the low literacy levels in deaf people [3]. Nowadays most written languages are speech-based alphabetic scripts, in which speech sound units (phonemes) are represented by visual symbols (letters, or graphemes). Learning the correspondences between letters and speech sounds of a language is therefore a crucial step in reading acquisition, failure of which is thought to account for reading problems in developmental dyslexia [4]. However, in the normal situation, letter-speech sound associations are learned and used with high efficiency. At least 90% of school children learn the letter-sound correspondences without exceptional effort with a few months [5], which are a remarkable achievement, since our brains are not phylogenetically adapted to the requirements for acquiring written language.

Associations between sensory events in different modalities can either be defined by natural relations (e.g., the shape and sound of a natural object) or by more artificial relations. In contrast to the culturally defined associations between letters and speech sounds [6], lip reading is based on naturally developed associations of speech with visual information [7]. Therefore, it seems a plausible assumption that the perception of speech and the inherently linked lip movements (hereafter referred to as *audiovisual speech*) emerged simultaneously during evolution, shaping the brain for integrating this audiovisual information. The present study was thus conducted the electroencephalography (EEG) experiments to characterize the influence of visual orthography on the most robust auditory event-related potentials (ERPs) and focused the analysis on systematic variation of the auditory ERPs as a function of visual orthography information. The subjects received auditory, visual, and audiovisual letters and were required to identify them, regardless of stimulus modality. Audiovisual letters included matching letters, in which the auditory and visual stimulus corresponded to each other based on previous experience, and nonmatching (randomly paired) letters. Meaningless auditory, visual, and audiovisual control stimuli were presented as well. The brain activations were detected with EEG, which is well suited for noninvasive identification of cortical activity and its accurate temporal dynamics.

MATERIALS AND METHODS

Subjects

Subjects were healthy and had normal hearing and vision (self reported). Fourteen adult, native speakers of Korean (7 males; 7 females), were participated in the ERPs experiment. Subjects were closely matched with respect to age (mean = 25.2, SD = 3.2) and years of formal education (mean = 18.2, SD = 2.2). No subject has any previous exposure to Chinese or for that matter any other tone language. None had any musical training within the past five years. All subjects were paid for their participation. They gave informed consent in compliance with a protocol before participation.

Stimuli

Stimuli consisted of a set of four Mandarin Chinese words that are distinguished minimally by tonal contour (*pinyin* Roman transliteration): yi^1 'clothing' [T1]; yi^2 'aunt' [T2]; yi^3 'chair' [T3]; yi^4 'easy' [T4]. Only three of the three Mandarin Chinese tones (T1, T2, T3) were chosen for presentation in a oddball paradigm. This limitation restricted EEG recording time to 90 mins, thus minimizing the risk of subject fatigue. The experiment consisted of an oddball condition. The duration of the stimuli were 300 ms. The audiovisual experiment included four stimuli: congruent /yi^1/ (acoustic /yi^1/ + visual /yi^1/), congruent /yi^2/ (acoustic /yi^2/ + visual /yi^2/), incongruent /yi^1/ (acoustic /yi^1/ + visual /yi^2/) and congruent /yi^4/ (acoustic /yi^4/ + visual /yi^4/). The auditory and visual experiments included only the acoustic and the visual parts of these stimuli, respectively (see figure 1 – figure 4).

Stimulus presentation

Stimulus sequences were presented to the subjects with STIM2 software. The stimulus onset asynchrony was 1300 ms (from acoustic/visual speech onset to onset). Stimulus sequences consisted of frequent (probability (P) = 0.60) congruent /yi^1/ stimuli and congruent (P = 0.15) and incongruent (P = 0.15) /yi^2/ stimuli. Congruent /yi^4/ stimuli were presented as target (P = 0.10) to be able to check that subjects were attending the stimuli. Randomized stimulus sequences were presented consisting of equiprobable auditory stimuli, visual stimuli, and audiovisual stimuli (a simultaneous combination of auditory and visual). Acoustic stimuli were delivered binaurally to the subjects through plastic tubes and earpieces. Sound density was adjusted to be 85 dB above the subject's hearing threshold (defined for the audiovisual stimulus sequence). Visual stimuli were presented on the computer screen. In the visual experiment, acoustic stimuli were not presented, but it was similar to audiovisual experiment in all other respects. Frequent (P = 0.60) / yi^1/ stimuli will be called visual standards and infrequent (P = 0.30) /yi^2/ stimuli visual deviants and /yi^4/ stimuli visual target (see figure 1 – figure 4). In an initial practice run, the task difficulty (i.e. target discriminability) was individually adjusted to about 75% correct responses for both auditory and visual target stimuli.

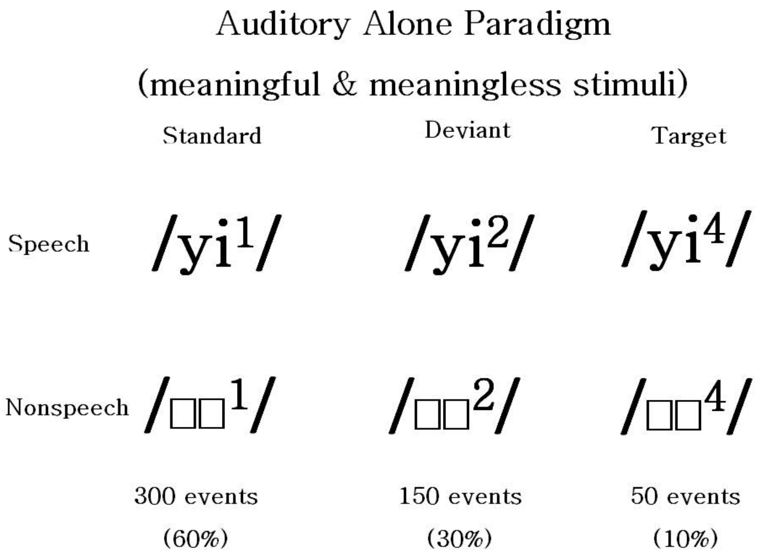

Figure 1: The auditory alone experiment included three stimuli for meaningful and three stimuli for meaningless, respectively. Standard: /yi^1/ (acoustic /yi^1/ and / □□^{1}/); Deviant: /yi^2/ (acoustic /yi^2/ and / □□^{2}/); and Target: /yi^4/ (acoustic /yi^4/ and / □□^{4}/).

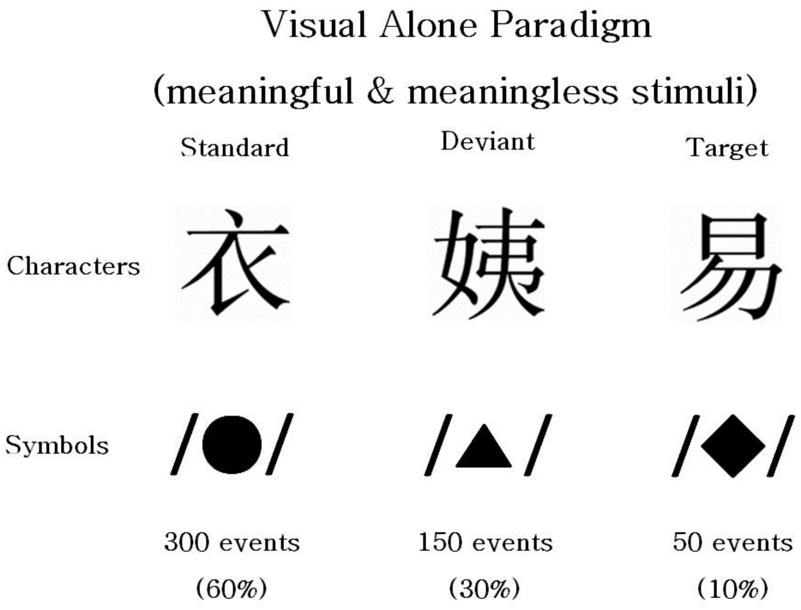

Figure 2: The visual alone experiment included three stimuli for meaningful and three stimuli for meaningless, respectively. Standard: /yi^1/ (Chinese character /yi^1/ and symbol /●/); Deviant: /yi^2/ (Chinese character /yi^2/ and /▲/); and Target: /yi^4/ (Chinese character /yi^4/ and symbol /♦/).

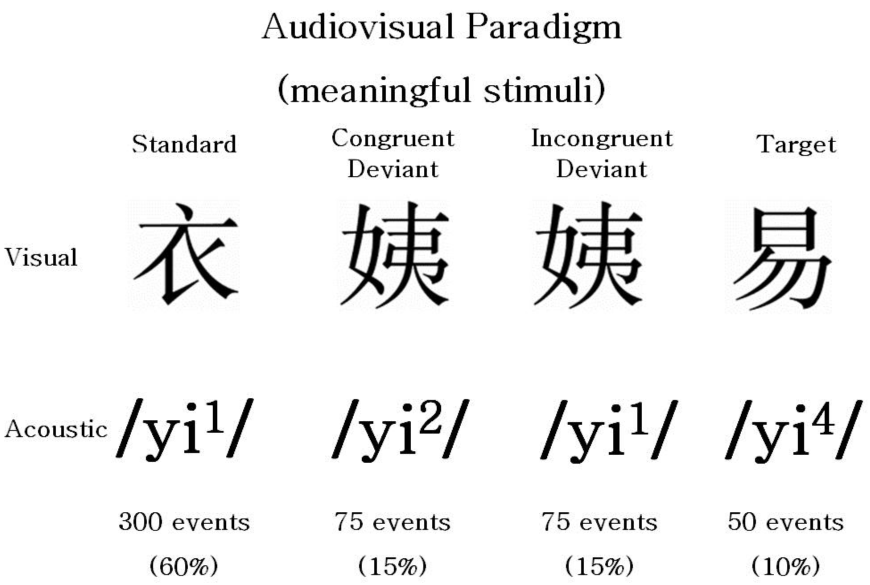

Figure 3: The audiovisual experiment included four stimuli. Standard: congruent /yi^1/ (acoustic /yi^1/ + visual/ yi^1/); Deviant: congruent /yi^2/ (acoustic /yi^2/ + visual /yi^2/); Deviant: incongruent /yi^1/ (acoustic /yi^1/ + visual /yi^2/); and Target: congruent /yi^4/ (acoustic /yi^4/ + visual /yi^4/).

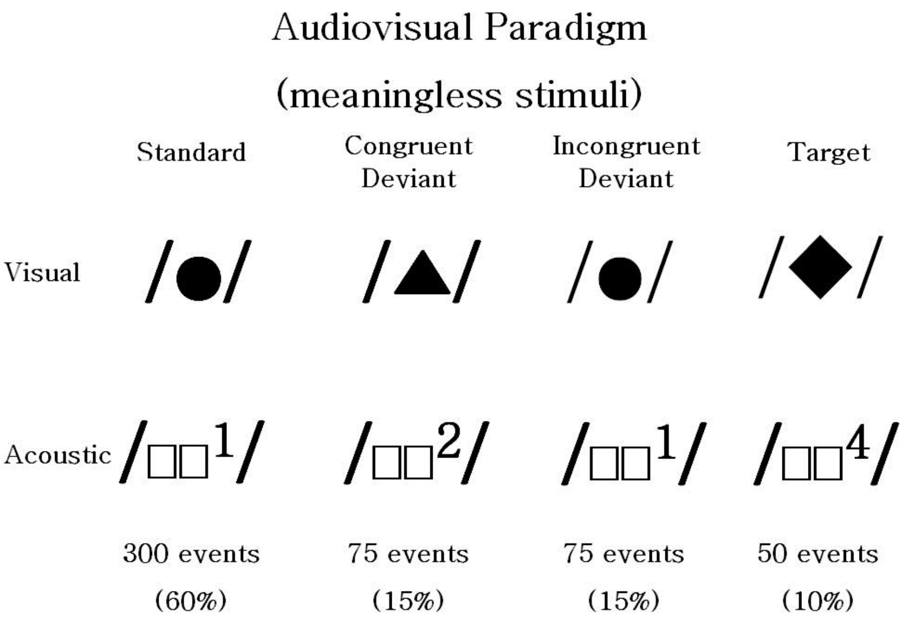

Figure 4: The audiovisual experiment included four stimuli. Standard: congruent /yi^1/ (acoustic / □□^{1}/ + visual /●/); Deviant: congruent /yi^2/ (acoustic / □□^{2}/ + visual /▲/); Deviant: incongruent /yi^1/ (acoustic / □□^{1}/ + visual /●/); and Target: congruent / □□^{4}/ (acoustic /yi^4/ + visual /♦/).

Experiment

Each experiment consisted of 2 blocks and each block had 300 trials. There were 6 blocks of all experiments. Every stimulus was presented with 300 ms exposure duration and inter-stimulus interval (ISI) was 1000ms in every condition. Subjects sat in an electrically shielded and soundproofed room with the response buttons under their hands. The subject had to press the button on the response pad when the target was presented and ignore any other types of stimuli. Prior to the experimental session, a practice block was administrated to ensure that the subjects understood the task. In the audiovisual condition, the subject was instructed to pay attention to the letters (orthography) and ignore the auditory stimuli.

Event-Related Potential (ERP) Recordings

EEG data were collected in an electrically and acoustically shield room. EEG was recorded with a Quick-Cap equipped with 64 channels according to the international 10-20 system using Scan system (Scan 4.3, Neurosoft, Inc. Sterling, USA) (see figure 5). Reference electrode was at mastoids. The signals were bandpass filtered at 0.05-100 Hz and digitized at 1000 Hz. The impedance of the electrode was below 5 kΩ. Eye movements were monitored with two EOG electrodes. Four electrodes monitored horizontal and vertical eye movements for off-line artifact rejection. Vertical and horizontal EOG was recorded by electrodes situated above and below the left eye, and on the outer canthi of both eyes, respectively. Epochs with EEG or EOG with a large (>100 µV) amplitude were automatically rejected. The artifact-free epochs were filtered at 0.1-15 Hz, baseline corrected and averaged.

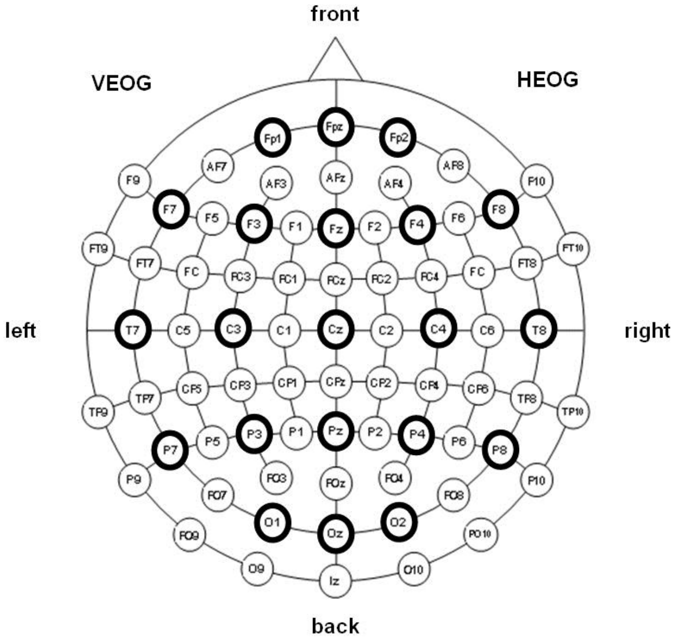

Figure 5: The 64-Channel Electrode Montage: ERPs from 21 channels (circle) were selectively analyzed in this experiment.

Data analysis

After the data recordings, the EEG was segmented into 1000 ms epochs, including the 100 ms pre-stimulus period. The baseline was corrected separately for each channel according to the mean amplitude of the EEG over the 100 ms period that preceded stimulus onset. The EEG epochs contained amplitudes exceeding ±100 µV at any EEG channels were automatically excluded from the averaging. The epoch was separately averaged for the standard, deviant, and the target stimulus. The average waveforms obtained from the standard, deviant and target stimuli were digitally filtered by a 0.1 - 15 Hz band-pass filter and finally baseline-corrected. The N1 that was elicited at approximately 100 ms after the onset of auditory stimulus, was visually inspected from waveform of standard and deviant stimulus. Cross-modal interaction was investigated by subtracting the ERPs to the auditory (A) and the visual (V) stimuli alone from the ERP to the combined audiovisual (AV) stimuli (i.e. interaction = AV - (A+V) and was identified as the peak voltage between 100-250 ms after stimulus onset in the subtracted waveform. The amplitude of the difference waveform was expressed in microvolt and its latency in milliseconds. Only ERPs to the standard stimuli were included in this analysis. By using a peak-detection algorithm, the negative peak was identified in the AV - (A+V) difference waveform between 100 – 250 ms. For statistical testing two-tailed *t*-tests were carried out comparing mean amplitudes within specified time windows that included the peak against the -100 to 0 ms pre-stimulus base line.

Statistical analysis

Statistical analysis was performed on the Global Field Power (GFP) area of 21 electrodes sites within the time range of difference waveform of cross-modal interaction (100-250 ms). Two conditions were speech and non-speech sounds. Five sites were prefrontal line, frontal line, central line, parietal line, and occipital line, respectively. ERP was analyzed with two-way ANOVAs with a repeated measure (condition x electrode site). Four electrodes sites such as prefrontal line (FP1, FPz, Fp2), frontal line (F7, F3, Fz, F4, F8), central line (T7, C3, Cz, C4, T8), parietal midline (P7, P3, Pz, P4, P8), and occipital line (O1, Oz, O2) sites were used.

RESULT

The grand-average ERPs audiovisual (AV), summed of auditory and visual (A+V), and audiovisual integration (AV-integration) waveforms for the two conditions (meaningful and meaningless) at the 12 electrode sites (Left: F3, Fz, F4, C3, Cz, C4; Right: P3, Pz, P4, O1, Oz, O2) are shown in Figures 6-7. The waveforms showed a typical morphology that indicated audiovisual P1, N1, P2, and P3 components. Irrespective of group, the AV integration elicited from AV – (A + V). It can be seen at frontal sites that the stimulus-triggered ERPs were preceded by a slow ramping positivity that apparently began before stimulus presentation, which was paralleled by a slow negativity over parietal and occipital sites.

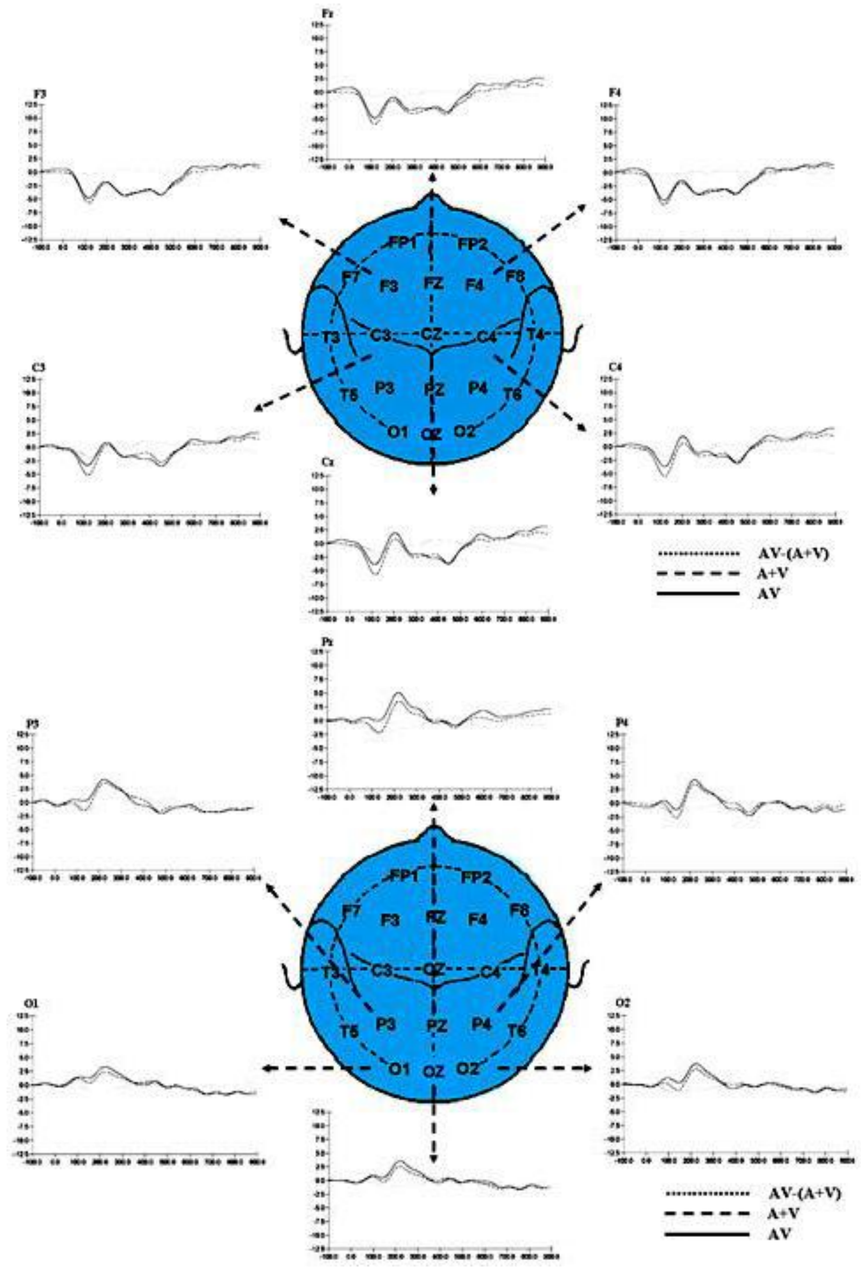

Figure 6: Superimposition of grand-average ERPs to bimodal (AV) stimuli, the algebraic sum of ERPs to unimodal auditory (A) and visual (V) stimuli, and the difference between the two waveforms (i.e. AV-(A+V)) at a subset of anterior (top) and posterior (bottom) electrodes in meaningful stimuli perception.

The thick traces in Figures 6-7 show the cross-modal interaction waveform [AV-(A+V)] obtained by subtracting the ERPs to the auditory (A) and visual (V) unimodal stimuli from the ERP to the bimodal audiovisual (AV) stimuli. It can be seen that the bimodal response is not simply the linear sum of separately recorded unimodal activity. A two-way [condition x electrode location] repeated measures ANOVA conducted on the mean amplitude yielded main effects of group ($F_{1,23} = 7.31$, $p = 0.01$) and condition ($F_{1,23} = 9.63$, $p < 0.01$). No main effect of electrode location was found ($F_{1,26} = 1.44$, $p = 0.282$). Subjects showed a larger mean amplitude for the both condition ($F_{1,26} = 9.73$, $p < 0.01$). Comparing conditions (Meaningful vs. Meaningless), the mean amplitude response was significantly less in meaningful stimuli condition than in the meaningless stimuli condition ($F_{1,26} = 11.41$, $p < 0.01$) (see Figure 8). In addition, the peak later for the meaningful stimuli (116 ms) relative to the meaningless stimuli (89 ms). A repeated measures two-way ANOVA [condition x electrode location] conducted on the peak latency measure yielded a significant main effect of condition ($F_{1,23} = 16.40$, $p < 0.01$), indicating that the peak of the integration occurred later in time for the meaningful stimuli relative to the meaningless stimuli. No other main effects or interaction effects reached significance (see Figure 8).

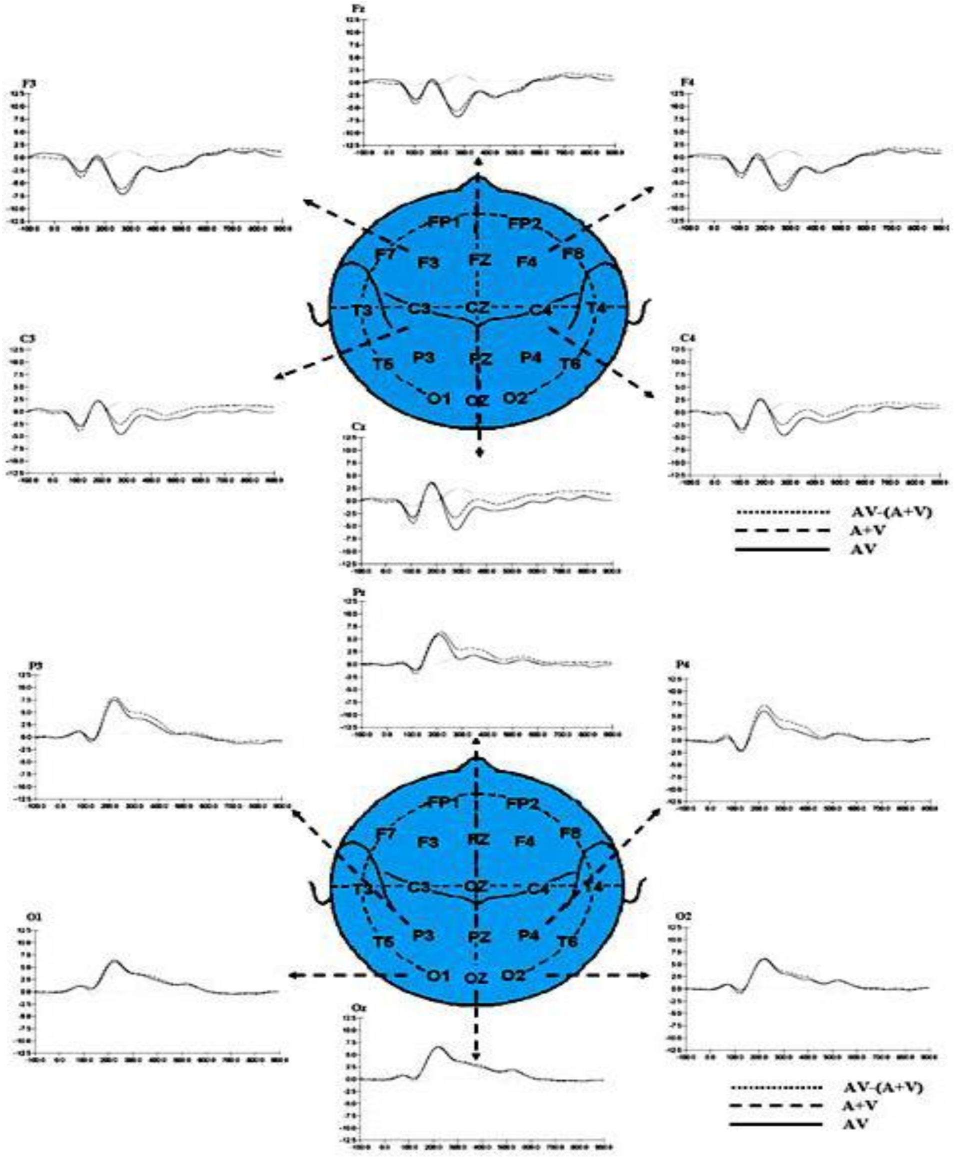

Figure 7: Superimposition of grand-average ERPs to bimodal (AV) stimuli, the algebraic sum of ERPs to unimodal auditory (A) and visual (V) stimuli, and the difference between the two waveforms (i.e. AV-(A+V)) at a subset of anterior (top) and posterior (bottom) electrodes in meaningless stimuli perception.

DISCUSSION

The present study was able to find evidence of both non-phonetic and phonetic audiovisual interactions in the ERPs to the same AV stimuli. The differences in the sum of the ERPs to the unimodal A and V stimuli and in ERPs to AV stimuli indicated interactions presumably based on temporal of the A and V components of the AV stimuli. In addition, the differences in the ERPs to the meaningful and meaningless of AV stimuli probably reflect multisensory interactions in phonetic processing. When acoustic and visual phonemes were meaningful, they formed a natural multisensory interaction stimulus.

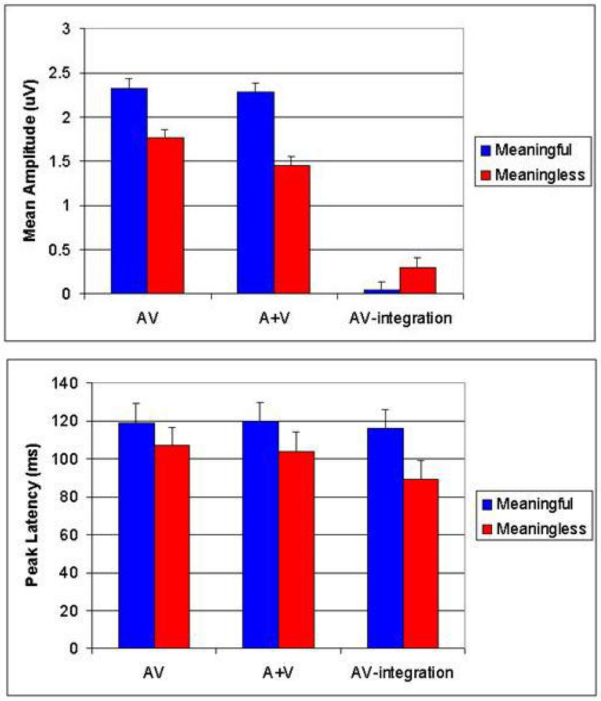

Figure 8: Mean peak amplitude (top) and latency (bottom) values are displayed for the two stimuli groups (meaningful, meaningless) per experimental condition (AV, A+V, and AV integration) as measured from all electrodes (GFP).

Additionally, the present results provide evidence that waveform deflection can contribute to the bimodal minus unimodal difference waveform [AV - (A+V)] that is often taken as an index of cross-modal interactions in neural processing. Superimposed upon these deflections in the AV-(A+V) difference wave were waves that appeared to reflected true cross-modal interactions. This interaction thus appears to take place in cortical areas of the parieto-occipital region. A similar auditory-visual interaction was observed at occipital sites by Giard and Peronnet [8] at 155-220 ms, which they interpreted as a modulation of the visual evoked N1 wave. This effect does indeed appear to represent an influence of auditory input on processing in a predominantly visual cortical area. The second major deflection indicative of cross-modal interaction peaked at 220-250 ms and could be accounted for by a dipole pair in anterior temporal peri-sylvian cortex. This effect might represent an interaction in auditory association cortex or in Polymodal cortex of the superior temporal plane [9].

A neural mechanism for the integration of audiovisual speech has been suggested by Calvert and colleagues [9, 10] and supported by other neuroimaging findings on audiovisual speech perception [11-13] and lip reading [7, 14, 15]. Results of these studies suggest that the perceptual gain experienced when perceiving multimodal speech is accomplished by enhancement of the neural activity in the relevant sensory cortices. The left posterior superior temporal sulcus (STS) has been advanced as the heteromodal site that integrates visual and auditory speech information and modulates the modality-specific cortices by back projections [9,10]. Modality-specific regions involved in this mechanism are the visual motion processing area V5 and auditory association areas in superior temporal cortex. In addition to this interplay between STS and sensory cortices, frontal and parietal regions seem to be involved, although activation of these regions is less consistent between the different studies. Interestingly, the involvement of the left posterior STS in the integration of auditory and visual nonlinguistic information has also been reported recently [16, 17]. These results suggest that the STS have a more general role in the integration of cross-modal identity information. According to the hypothesis by Calvert *et al.* [9], unimodal speech signals are integrated in STS and fed back onto primary auditory areas. This mechanism predicts activation of auditory cortices during visual [14] and enhanced activation during audiovisual speech processing [12, 18, 19]. However, the AV interactions in the left auditory cortex might precede those in the right STS [20] and the audiovisual responses measured with electroencephalogram (EEG) and magnetoencephalogram (MEG) are often smaller during audiovisual stimulation than the sum of responses during unimodal stimulation [20-23].

According to Calvert *et al.* [15] and Callan *et al.* [19], both integration in STS and the internal articulatory simulation of the intended speech act of the (visually) observed speaker facilitate auditory speech perception through back-projections to auditory cortical areas. Articulatory simulation of the visual speech input would have a secondary role in audiovisual speech perception and is used to facilitate primary acoustic-phonetic processing especially in sub-optimal conditions [15, 19]. Despite the similarities, this account is different from the one presented above. The important difference is that in these models the visual speech input is assumed to be processed independently of the auditory input in the speech motor regions. Visual influence on auditory processing is achieved without convergence of A and V inputs into motor representations.

CONCLUSION

The present study demonstrates the sensory-specific and heteromodal cortical regions which are involved in the AV integration process at separate latencies and are sensitive to different features of AV speech stimuli. The auditory and visual speech interacts in the auditory cortical regions early on in the processing hierarchy. The audiovisual interaction following elementary within-modality discrimination processes imply the attention-related rechecking of the outcome of within-modality analyses. The processing of a feature, hierarchically dependent on another feature and support the view on the neurocognitive mechanisms of audiovisual speech perception which emphasizes the involvement of multiple, hierarchically organized and mutually interacting brain mechanisms.

REFERENCES

[1] National Reading Council. Preventing Reading Difficulties in Young Children, Washington, DC: National Academy Press, 1998.

[2] Liberman AM. The relation of speech to reading and writing. In Frost R, Katz L. (eds.). Orthography, Phonology, Morphology and Meaning, Amsterdam: Elsevier Science Publishes B.V., 1992, pp. 167-178.

[3] Perfetti CA, Sandak R. J. Deaf Stud. Deaf Edu 2000; 5: 32-50

[4] Frith U. Beneath the surface of developmental dyslexia. In Patterson K E, Marshall J C, Coltheart M. (eds.). Surface Dyslexia, London: Routledge & Kegan-Paul, 1985, pp. 301-330.

[5] Blomert L, Dyslexie: Stand van Zaken (Dyslexia: State of Affairs in The Netherlands). Report for the Dutch Ministry of Health. In Dyslexie naar een vergoedingsregeling, R. Reij (ed.). Amstelveen: Dutch Health Care Insurance Board, 2002.

[6] Raij T, Uutela K, Hari R. Neuron 2000; 28: 617-627.

[7] Paulesu E, Perani D, Blasi V, Silani G, Borghese NA, De Giovanni U, Sensolo S, Fazio F. J Neurophysiol 2003; 90: 2005-2013.

[8] Giard MH, Peronnet F. J Cognitive Neurosci 1999; 1: 473-490.

[9] Calvert GA, Campbell R, Brammer M. J. Curr Biol 2000; 10: 649-657.

[10] Calvert GA, Brammer MJ, Bullmore ET, Campbell R, Iversen SD, David AS. Neuroreport. 1999; 10: 2619-2623.

[11] Sams M, Aulanko R, Hämäläinen M, Haru R, Lounasmaa, OV, Lu S-T. Neurosci. Lett 1991; 127: 141-145.

[12] Sekiyama K, Kanno I, Miura S, Sugita Y. Neurosci Res 2003; 47: 277-287.

[13] Wright TM, Pelphrey KA, Allison T, McKeown, MJ, McCarthy G. Cereb Cortex 2003; 13: 1034-1043.

[14] Calvert GA, Bullmore ET, Brammer MJ, Campbell R, Williams SC, McGuire PK. Science 1997; 276: 593-6.

[15] Calvert GA, Campbell R. J. Cognitive Neurosci 2003; 15: 57-71.

[16] Beauchamp M, Lee K, Argall B, Martin A. Neuron 2004; 41: 809-823.

[17] Calvert GA, Hansen PC, Iversen SD, Braummer MJ. NeuroImage 2001; 14: 427-438.

[18] Callan DE, Jones JA, Munhall KG, Callan AM, Kroos C, Vatikiotis-Bateson E. Neuroreport 2003; 14: 2213-2217.

[19] Callan DE, Jones JA, Munhall K, Kroos C, Callan AM, Vatikiotis-Bateson E. J Cognitive Neurosci., 2004; 16: 805-16.

[20] Möttönen R, Schurmann M, Sams M, Neurosci Lett 2004; 363: 112-5.

[21] Besle J, Fort A, Delpuech C, Giard MH. Eur. J Neurosci 2004; 20: 2225-34.

[22] Klucharev V, Möttönen R, Sams M. Brain Res Cog Brain Res 2003; 18: 65-75.

[23] van Wassenhove V, Grant, KW, Poeppel D. P. Natl Acad Sci USA 2005; 102: 1181-6.