Electrophysiological evidence of illusory audiovisual speech percept in human infants

Elena Kushnerenko*†‡, Tuomas Teinonen*§, Agnes Volein*, and Gergely Csibra*

*Centre for Brain and Cognitive Development, School of Psychology, Birkbeck University of London, Malet Street, London WC1E 7HX, United Kingdom; †Institute for Research in Child Development, School of Psychology, University of East London, London E15 4LZ, United Kingdom; and §Cognitive Brain Research Unit, Department of Psychology, University of Helsinki, P.O. Box 9, FI-00014 Helsinki, Finland

Edited by Dale Purves, Duke University Medical Center, Durham, NC, and approved June 16, 2008 (received for review May 2, 2008)

How effortlessly and quickly infants acquire their native language remains one of the most intriguing questions of human development. Our study extends this question into the audiovisual domain, taking into consideration visual speech cues, which were recently shown to have more importance for young infants than previously anticipated [Weikum WM, Vouloumanos A, Navarra J, Soto-Faraco S, Sebastián-Gallés N, Werker JF (2007) *Science* 316:1159]. A particularly interesting phenomenon of audiovisual speech perception is the McGurk effect [McGurk H, MacDonald J (1976) *Nature* 264:746–748], an illusory speech percept resulting from integration of incongruent auditory and visual speech cues. For some phonemes, the human brain does not detect the mismatch between conflicting auditory and visual cues but automatically assimilates them into the closest legal phoneme, sometimes different from both auditory and visual ones. Measuring event-related brain potentials in 5-month-old infants, we demonstrate differential brain responses when conflicting auditory and visual speech cues can be integrated and when they cannot be fused into a single percept. This finding reveals a surprisingly early ability to perceive speech cross-modally and highlights the role of visual speech experience during early postnatal development in learning of the phonemes and phonotactics of the native language.

audiovisual integration | event-related potentials (ERP) | mismatch negativity (MMN) | speech perception

Human infants acquire language by intensively analyzing the distributional patterns of the speech environment (1, 2). There is, however, something that they never encounter in everyday life: speaking lips that do not match the perceived sound. Speech perception is multimodal: Seeing lip movements influences the perception of auditory information. When a mismatch occurs between auditory and visual speech, people frequently report hearing phonemes different from the stimulus in either modality. For example, when an auditory /ba/ syllable is dubbed onto a visual /ga/ syllable, the most common resulting percept is /da/, representing an illusory fusion between the actual stimuli in the two modalities (3). However, not all combinations of visual and auditory syllables form a coherent speech percept: an auditory /ga/ presented with a visual /ba/ will likely to be detected as an audiovisual mismatch or will be heard as an illusory combination of the two syllables (/bga/). This combination effect has been found less reliably than the fusion effect in adults (54% and 98%, respectively), and it was very rare in 3- to 5-year-old children (19% occurrence of the combination effect compared to 81% occurrence of the fusion effect) (3). The discrepancy between the two subtypes of audiovisual integration may be explained by the resulting illusory percepts: whereas the fusion effect results in a legal syllable (e.g., /da/), the illusory combination percept /bga/ contains a consonant cluster /bg/ that is phonotactically illegal at the beginning of words in English and many other languages.

Human infants are exposed to talking faces from the first minutes of life. The concurrent audiovisual stimulation that they experience results in rapid formation of cross-modal neural phoneme representations. The visual component of speech may also contribute to the learning of auditory phoneme categories (4). Behavioral studies have shown that 2- to 4-month-old infants can already match heard vowels with the appropriate lip movements (5, 6). Recently, it was reported that visual speech information alone is sufficient for language discrimination in 4- to 6-month-old infants (7). Illusory fusion, however, requires not just matching but integration and is characterized by the absence of detection of the auditory–visual mismatch. Two studies have indicated that infants may perceive illusory fusion at 4 months of age (8, 9). However, the conclusiveness of these results was disputed and it was suggested that behavioral measures alone would be insufficient for resolving this issue (10). Measuring the neural correlates of audiovisual integration could clarify whether infants detect the mismatch between conflicting stimuli or process them as a unified percept.

Event-related brain potentials (ERPs) and event-related oscillations (EROs) are generated by neural processes that are time-locked to significant events, such as stimulus onset, and can be reliably recorded from infants (11). In adults, an ERP component that is predominantly elicited for changes in the auditory modality [the mismatch negativity, MMN, (12)] can also be observed when only visual components of audiovisual speech changes, with no real acoustic variation (13, 14). However, because the recording of MMN requires extensive presentation of stimuli with differing frequencies, we chose an alternative method, less taxing for young infants' attention, for investigating audiovisual integration in the early months of life.

We recorded the electrical brain activity of 17 5-month-old infants while they were watching, and listening to, audiovisual (AV) syllables of four types in random order with equal probability [see supporting information (SI) Movie S1, Movie S2, Movie S3, and Movie S4]. Two AV stimuli represented canonical /ba/ and /ga/ syllables, whereas the other two were generated by crossing the auditory and visual components of these stimuli. One of these mixed stimuli, which included a visual /ga/ and an auditory /ba/ (VgAb) component would typically result in illusory fusion, whereas the opposite mixture (VbAg) may generate illusory combination or the perception of incongruency. We predicted that if infants were subject to the illusory fusion effect, their brain responses to the VgAb stimulus would not differ from those of ordinary /ba/ and /ga/ syllables, whereas they should show evidence of detecting the audiovisual mismatch in the opposite VbAg stimulus. This stimulus was expected to be processed as an "odd" syllable, violating long-term memory

Author contributions: E.K. and G.C. designed research; E.K., T.T., and A.V. performed research; E.K. and T.T. analyzed data; and E.K. and G.C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

‡To whom correspondence should be addressed at: Institute for Research in Child Development, School of Psychology, University of East London, Romford Road, London E15 4LZ, United Kingdom. E-mail: e.kushnerenko@gmail.com.

This article contains supporting information online at www.pnas.org/cgi/content/full/0804275105/DCSupplemental.

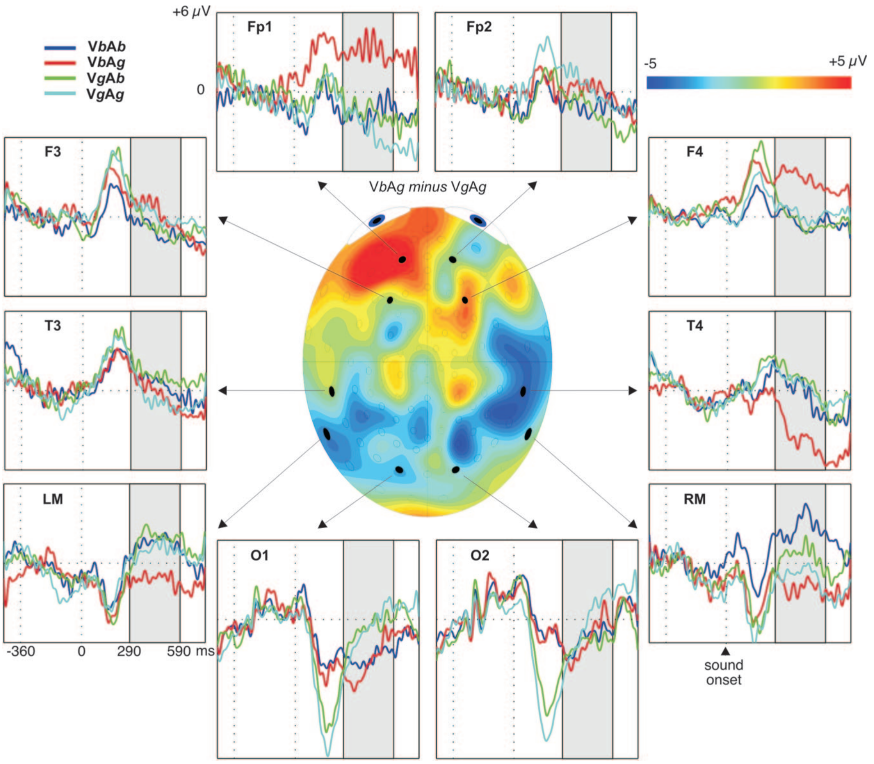

Fig. 1. Grand-averaged responses to the AV pairs VbAb (dark blue), VbAg (red), VgAb (green), and VgAg (light blue) time-locked to the sound onset at 0 ms. Selected channels are shown according to 10–20 system (LM and RM refer to the left and right mastoid, respectively). The highlighted area corresponds to a time window within which the ERP to the VbAg stimulus significantly deviated from the others. The topographic map in the middle represents the voltage difference between responses to VbAg and VgAg pairs within the highlighted window.

traces for allowable phonetic combinations. In other words, if infants successfully integrate the VgAb mixture, which generates illusory fusion in adults, into a single percept, only the phonotactically illegal VbAg is expected to generate a mismatch response. If, however, infants detect both kinds of mismatch, one would expect to see brain activation different from those of the canonical /ba/ and /ga/ syllables after both VbAg and VgAb stimuli.

Results

We analyzed the ERP responses to the visual stimuli first. We found an early and robust separation of occipital ERPs to the stimuli that contained visual /ga/ and the ones that contained visual /ba/. A two-way ANOVA on the amplitude within the 190- to 290-ms interval (measured from the onset of the auditory component) revealed a strong main effect of AV stimulus: F (3, 48) = 8.23, P = 0.0002). Post hoc least-significant difference (LSD) tests showed that the negative peak over occipital areas (see the recording from O1 and O2 on Fig. 1) was larger in response to AV stimuli composed with visual /ga/ than those composed with visual /ba/ (P = 0.0007–0.001).

The effect of the auditory–visual interaction could be detected on the frontal and temporal leads (Fig. 1). The ERP response to the VbAg stimulus was more positive over frontal areas (Fp1, F4) and more negative over temporal areas (LM, T4) than the ERPs to any other stimulus type. The ERPs started to deviate from the responses to the other AV stimuli at ≈290 ms after the sound onset, peaking at ≈360 ms and lasting beyond the epoch of analysis. Three-way (stimulus × location × hemisphere) ANOVAs on the ERP amplitudes in consecutive 100-ms-wide time windows revealed significant interactions between stimulus and location from 290 to 590 ms (F (6, 96) = 3.15 to 3.96; P = 0.001–0.007). Post hoc LSD tests showed that only the VbAg stimulus differed from all others, being more positive at frontal leads (P = 0.002–0.008) and more negative at temporal areas (P = 0.002–0.004). This inversion of polarity (Fig. 1, see also Fig. S1) suggests that the source of this effect lay in bilateral anterior temporal areas, possibly in the auditory cortex.

Discussion

In the current study, we used event-related potentials to examine 5-month-old infants' neural processing of conflicting audiovisual speech stimuli, known to be perceived by adults as "speech illusions." The data revealed that only the combination of a visual /ba/ and an auditory /ga/ was processed as mismatched audiovisual input, resulting in additional activation starting at ≈290 ms from sound onset over frontal and temporal areas.

An earlier effect was also observed for the difference in visual components of the AV stimuli: The occipital early negative response was larger in amplitude for the pairs composed with visual /ga/ than those with visual /ba/ (Fig. 1). This effect reflects the fact that articulating /ga/ involves earlier and faster opening of the mouth and, thus, seeing this stimulus generates an earlier and bigger response from the visual cortex than the lip movement corresponding to the /ba/ stimulus. There were no further modality-specific differences between the responses in subsequent time windows.

The additional frontal and temporal brain activation in response to the incongruent VbAg stimulus cannot be explained by auditory or visual processing alone, because congruent and incongruent AV pairs were presented with equal probability and composed of identical auditory and visual syllables. Instead, it reflects the detection of the mismatch between the auditory and visual components of the input. The preparatory movements for the articulation of the syllables started ≈260 ms for /ga/ and 280 ms for /ba/ before sound onset (averaged across speakers).

Therefore, the auditory input could have violated the stimulus anticipation generated by the leading visual input in both conflicting audiovisual combinations. However, the incongruent stimulus pair VgAb, which generates an illusory /da/ percept in adults, was not processed as a conflicting AV stimulus by the infants, as indicated by the lack of ERP difference when compared to the congruent /ba/ and /ga/ syllables. Thus, the infants in our experiment failed to detect the mismatch between the auditory and visual components of this AV stimulus, suggesting that they successfully integrated its components into a unified, and probably illusory, percept.

The perceptual outcome of audiovisual integration may depend on the ease of perceptual categorization of the visual stimulus (15). Place of articulation for /ba/ (lips pressing) restricts perceptual outcomes to /b/, /p/, or /m/ only, whereas articulating /ga/ (mouth opening) does not have such predictive power on the ensuing auditory event. Thus, the difference in processing the incongruent pairs VgAb (fusion effect) and VbAg (combination effect) could be due to the saliency of the visual /ba/ component, which disallowed its fusion with the auditory /ga/ input.

It is important to note that this effect may not be observable for all languages. Some languages, such as Japanese, provide less distinctive visual information, possibly explaining a weaker visual influence on AV integration in Japanese participants (16).

During the first year of life, broadly tuned sensory systems undergo the process of perceptual narrowing. Perceptual discrimination for stimuli predominant in the environment improves, whereas for stimuli not present in the environment, it declines (17). Accordingly, infants become better at discriminating native phonemes (18) and human faces (19), whereas they lose the ability of telling apart some nonnative phoneme contrasts and individuals of other species. They also become less sensitive to the intersensory match of non-species-specific faces and voices (20). Whether or not such perceptual tuning contributes to the development of audiovisual integration demonstrated in this study is a question for future research.

The timing (peaking ≈360 ms from sound onset) and scalp distribution (positive over frontal areas and negative over temporooccipital areas, see Fig. 1) of the ERP response to VbAg AV mismatch resembles that of the auditory mismatch response in infants (21–23). Even though the majority of the studies interpret the mismatch response in adults (MMN) as signaling the detection of occasional stimulus changes in a repetitive acoustic environment, one underlying mechanism has been suggested to reflect the activation of cortical cell assemblies forming the long-term memory traces for cognitive representations of sounds (24, 25). It is thus possible that the neural response we discovered indicates that the incongruent VbAg syllable failed to match with the infants' long-term memory traces for permissible phonemes or permissible AV relations learned by 5 months of age.

Although this study did not employ the traditional oddball paradigm used in previous research for eliciting mismatch and novelty responses in infants (11, 26), the mechanism underlying this AV ERP response may involve the detection of a violation of expectation. The visual information that precedes the auditory signal has been proposed to engage the speech-processing system by generating an internal prediction of the ensuing auditory signal (15). Van Wassenhove and colleagues (15) found reduced auditory evoked potentials to speech stimuli predicted by the visual input. Presumably, with the VbAg stimulus in our study, the leading /ba/ visual input triggered speech processing for /ba/ syllable, whereas the ensuing unexpected auditory input, which activated the neural representation of the /ga/ syllable, generated a mismatch response.

In summary, the audiovisual mismatch response recorded in our study indicates that, by the age of 5 months, infants have formed cross-modal neural representations of the syllables /ba/ and /ga/ and can generate anticipation for the oncoming auditory stimulus on the basis of visually perceived speech. This response is elicited only when infants are unable to integrate the audiovisual input. Intriguingly, the mismatch between auditory and visual input is not always detected, as in the McGurk illusion reported in adults. Our result suggests that, whenever possible, infants tend to assimilate the conflicting audiovisual input into a unified percept. This assimilation may serve as adaptive means to understanding the diversity of speakers with their individual differences in articulation.

Materials and Methods

Subjects. Seventeen healthy, full-term infants (10 girls, 7 boys) were tested between 20.5 and 23 weeks after birth (mean age 21.4 weeks, SD = 0.8 weeks). An additional seven infants were excluded from the analysis for excessive movements, fussiness, or bad electrode recording. The study was approved by the ethics committee of the School of Psychology of Birkbeck, University of London, and parents gave their written informed consent for their child's participation.

Stimuli. Video clips were recorded with three female native English speakers articulating /ba/ and /ga/ syllables (see Movie S1, Movie S2, Movie S3, and Movie S4). Sound onset was adjusted in each clip at 360 ms from stimulus onset, and the auditory syllable lasted for the following 280–320 ms. The video clips were rendered with a digitization rate of 25 frames per second, and the stereo soundtracks were digitized at 44.1 kHz with 16-bit resolution.

All AV stimuli contained nine frames (360 ms) of silent face before the sound onset, followed by the voiced part (seven or eight frames) and the mouth closing (two or three frames). The total duration of the AV stimuli was 760 ms. The mouth opening for the visual /ga/ stimulus started ≈260 ms before the sound onset (averaged across speakers), whereas it was simultaneous with the sound onset for the visual /ba/ stimulus. For visual /ba/, the lips started to press against each other ≈280 ms before the sound onset. Each AV stimulus started with lips fully closed and was followed immediately with the next AV stimulus, the stimulus onset asynchrony (SOA) being 760 ms.

For each of the three speakers, four categories of AV stimuli were presented with equal probability: congruent VbAb and VgAg and incongruent pairs (VgAb and VbAg). The incongruent pairs were created by using the original AV stimuli by dubbing the auditory /ba/ onto a visual /ga/ and vice versa. Therefore, each auditory and each visual syllable was presented with equal frequency during the task. The consonantal burst in the audio file was aligned with the consonantal burst of the original soundtrack of the video file.

The incongruent AV stimuli were presented to five native adult English speakers to test whether they produce illusory percepts. Four of them reported hearing /da/ or /ta/ for VgAb (fusion percept) and either /bga/ or mismatched audiovisual input for VbAg, and one adult reported only the auditory component in both situations.

Procedure. Syllables were presented in pseudorandom order, changing the speaker approximately every 40 s to keep the infant attentive. Videos were displayed on a 40 × 30-cm monitor on black background while the infant, sitting on a parent's lap, was watching them from 80-cm distance in an acoustically and electrically shielded booth. The faces on the monitor were approximately life size. Sounds were presented through two loudspeakers behind the screen at a 64- to 69-dB sound pressure level (noise level 30 dB). The recording time varied from 4 to 9 min, depending on the infant's attention to stimuli. The behavior of the infants was recorded on video tape.

EEG Recording and Analysis. The brain's electrical activity was recorded by using a 128-channel Geodesic Sensor Net (Electrical Geodesic) referenced to the vertex (27). The electrical potential was amplified, digitized at 500 Hz, and band-pass filtered from 0.1 to 200 Hz. The EEG was segmented into epochs starting 100 ms before and ending 1,100 ms after the AV stimulus onset. Channels contaminated by eye or motion artifacts were rejected manually, and trials with >20 bad channels were excluded. In addition, video recordings of the infants' behavior were coded off-line frame-by-frame, and trials during which the infant did not attend to the screen were excluded from further analyses. The first trial after each speaker change was also excluded. After artifact rejection, the average number of trials for an individual infant accepted for further analysis was 33.34 for /ba/, 32.41 for /ga/, 30.47 for VgAb, and 34.05 for VbAg.

The artifact-free trials were rereferenced to the average reference and then averaged for each infant within each condition. Averages were baseline-corrected for 200 ms before sound onset to minimize the effects of any ongoing processing from the preceding stimulus. Electroencephalographic and magnetoencephalographic studies in adults have reported AV interaction-related activation for speech as early as at 150–250 ms from sound onset (13, 15, 28, 29) and sometimes lasting up to 600 ms (30). Thus, we chose to calculate mean voltage in successive 100-ms time windows starting from 90 ms from sound onset up to 690 ms (i.e., 450–1,050 ms after AV stimulus onset).

For statistical analyses, we defined channel groups to represent the activation of the auditory cortex, which, according to earlier studies in adults and infants, can spread over frontocentral and temporal areas (bilateral frontal, central, and temporal channel groups, see Fig. S2). Activity from the visual areas was measured from bilateral occipital channel groups. Average amplitudes across channels within each channel group were entered into analyses of variance to evaluate the effects of stimulus conditions. Greenhouse–Geisser correction was used where applicable.

Two topographic maps of the voltage difference were created to control for possible differences caused purely by auditory or visual differences between the stimuli. Thus, in Fig. S1, we present the voltage difference between incongruent VbAg and congruent VgAg pairs differing only visually (Fig. S1*A*), and the voltage difference between VbAg pair and VbAb pair differing only in the auditory modality (Fig. S1*B*). The similarity between these two maps is consistent with the ANOVA results for this latency window (see *Results*). For γ-band oscillations in response to incongruent (VbAg) AG stimuli, see Fig. S3.

Data regarding γ-band event-related oscillations can be found in *SI Text*.

ACKNOWLEDGMENTS. We thank Dr. Fani Deligianni for assistance in programming. This study was supported by the Academy of Finland (Project 213672) and a Pathfinder grant (CALACEI) from the European Commission.

1. Maye J, Werker JF, Gerken L (2002) Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82:B101–B111.
2. Saffran JR, Aslin RN, Newport EL (1996) Statistical learning by 8-month-old infants. *Science* 274:1926–1928.
3. McGurk H, MacDonald J (1976) Hearing lips and seeing voices. *Nature* 264:746–748.
4. Teinonen T, Aslin, RN, Alku P, Csibra G (2008) Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition*, in press.
5. Kuhl PK, Meltzoff AN (1982) The bimodal perception of speech in infancy. *Science* 218:1138–1141.
6. Patterson ML, Werker JF (2003) Two-month-old infants match phonetic information in lips and voice. *Dev Sci* 6:191–196.
7. Weikum WM, *et al.* (2007) Visual language discrimination in infancy. *Science* 316:1159.
8. Burnham D, Dodd B (2004) Auditory-visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Dev Psychobiol* 45:204–220.
9. Rosenblum L, Schmuckler M, Johnson J (1997) The McGurk effect in infants. *Percept Psychophys* 59:347–357.
10. Desjardins RN, Werker JF (2004) Is the integration of heard and seen speech mandatory for infants? *Dev Psychobiol* 45:187–203.
11. Csibra G, Kushnerenko E, Grossman T in *Handbook of Developmental Cognitive Neuroscience*, eds Nelson CA, Luciana M, in press.
12. Näätänen R, Gaillard AWK, Mäntysalo S (1978) Early selective attention effect on evoked potential reinterpreted. *Acta Psychologica* 42:313–329.
13. Colin C, Radeau M, Soquet A, Demolin D, Colin FPD (2002) Mismatch negativity evoked by the McGurk–MacDonald effect: A phonetic representation within short-term memory. *Clin Neurophysiol* 113:495–506.
14. Saint-Amour D, De Sanctis P, Molholm S, Ritter W, Foxe JJ (2007) Seeing voices: High-density electrical mapping and source-analysis of the multisensory mismatch negativity evoked during the McGurk illusion. *Neuropsychologia* 45:587–597.
15. van Wassenhove V, Grant KW, Poeppel D (2005) Visual speech speeds up the neural processing of auditory speech. *Proc Natl Acad Sci USA* 102:1181–1186.
16. Sekiyama K, Burnham D (2008) Impact of language on development of auditory-visual speech perception. *Dev Sci* 11:306–320.
17. Scott LS, Pascalis O, Nelson CA (2007) A domain-general theory of perceptual development. *Curr Dir Psychol Sci* 16:197–201.
18. Werker JF, Tees RC (1984) Cross-language speech perception : Evidence for perceptual reorganization during the first year of life. *Infant Behav Dev* 7:49–63.
19. Pascalis O, de Haan M, Nelson CA (2002) Is face processing species-specific during the first year of life? *Science* 296:1321–1323.
20. Lewkowicz DJ, Ghazanfar AA (2006) The decline of cross-species intersensory perception in human infants. *Proc Natl Acad Sci USA* 103:6771–6774.
21. Dehaene-Lambertz G, Baillet S (1998) A phonological representation in the infant brain. *NeuroReport* 9:1885–1888.
22. Dehaene-Lambertz G, Dehaene S (1994) Speed and cerebral correlates of syllable discrimination in infants. *Nature* 28:293–294.
23. Dehaene-Lambertz G, Pena M (2001) Electrophysiological evidence for automatic phonetic processing in neonates. *NeuroReport* 12:3155–3158.
24. Näätänen R, Tervaniemi M, Sussman E, Paavilainen P, Winkler I (2001) ''Primitive intelligence'' in the auditory cortex. *Trends Neurosci* 24:283–288.
25. Pulvermuller F, Shtyrov Y (2006) Language outside the focus of attention: The mismatch negativity as a tool for studying higher cognitive processes. *Prog Neurobiol* 79:49–71.
26. Kushnerenko E, *et al.* (2007) Processing acoustic change and novelty in newborn infants. *Eur J Neurosci* 26:265–274.
27. Tucker DM (1993) Spatial sampling of head electrical fields: The geodesic sensor net. *Electroencephalogr Clin Neurophysiol* 87:154–163.
28. Möttönen R, Krause CM, Tiippana K, Sams M (2002) Processing of changes in visual speech in the human auditory cortex. *Brain Res Cogn Brain Res* 13:417–425.
29. Sams M, *et al.* (1991) Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex. *Neurosci Lett* 127:141–145.
30. Möttönen R, Schurmann M, Sams M (2004) Time course of multisensory interactions during audiovisual speech perception in humans: A magnetoencephalographic study. *Neurosci Lett* 363:112–115.