Developmental Science

Developmental Science (2013), pp 1–15 DOI: 10.1111/desc.12098

PAPER

Audio-visual speech perception: a developmental ERP investigation

Victoria C.P. Knowland,[1,2] Evelyne Mercure,[3] Annette Karmiloff-Smith,[4] Fred Dick[4] and Michael S.C. Thomas[4]

1. *School of Health Sciences, City University, London, UK*
2. *Department of Psychological Sciences, Birkbeck College, London, UK*
3. *Institute of Cognitive Neuroscience, UCL, UK*
4. *Department of Psychological Sciences, Birkbeck College, London, UK*

Abstract

Being able to see a talking face confers a considerable advantage for speech perception in adulthood. However, behavioural data currently suggest that children fail to make full use of these available visual speech cues *until age 8 or 9. This is particularly surprising given the potential utility of multiple informational cues during language learning. We therefore explored this at the neural level. The event-related potential (ERP) technique has been used to assess the mechanisms of audio-visual speech perception in adults, with visual cues reliably modulating auditory ERP responses to speech. Previous work has shown congruence-dependent shortening of auditory N1/P2 latency and congruence-independent attenuation of amplitude in the presence of auditory and visual speech signals, compared to auditory alone. The aim of this study was to chart the development of these well-established modulatory effects over mid-to-late childhood. Experiment 1 employed an adult sample to validate a child-friendly stimulus set and paradigm by replicating previously observed effects of N1/P2 amplitude and latency modulation by visual speech cues; it also revealed greater attenuation of component amplitude given incongruent audio-visual stimuli, pointing to a new interpretation of the amplitude modulation effect. Experiment 2 used the same paradigm to map cross-sectional developmental change in these ERP responses between 6 and 11 years of age. The effect of amplitude modulation by visual cues emerged over development, while the effect of latency modulation was stable over the child sample. These data suggest that auditory ERP modulation by visual speech represents separable underlying cognitive processes, some of which show earlier maturation than others over the course of development.*

Research highlights

- The electrophysiological correlates of audio-visual speech perception show a course of gradual maturation over mid-to-late childhood.
- Electrophysiological data reveal that the speed of processing auditory speech is modulated by visual cues earlier in development then is suggested by behavioural data with children.
- In adults, the attenuation of auditory ERP component amplitude by visual speech cues is interpreted as an effect of cross-modal competition.
- It is suggested that the shortening of auditory ERP component latency by visual cues in adults may represent the prediction of both content and timing of the up-coming auditory speech signal.

Speech is multisensory

During face-to-face interaction the perception of speech is a multisensory process, with visual cues available from the talking face according a substantial benefit to adult listeners. Audio-visual speech perception has been fairly extensively studied in the adult population, yet little is understood about the extent to which, or how, children make use of these powerful cues when learning language. The aim of this study was to illuminate this matter through event-related potential (ERP) recordings with a developmental sample to establish how visual input modulates auditory processing over mid-to-late childhood.

Visual speech cues, that is movements of the lips, jaw, tongue and larynx, correlate closely with auditory output

Address for correspondence: Victoria C.P. Knowland, Department of Language and Communication Science, City University, Northampton Square, London EC1V 0HB, UK; e-mail: victoria.knowland.1@city.ac.uk

(Chandrasekaran, Trubanova, Stillittano, Caplier & Ghazanfar, 2009). Such cues are of particular benefit to adult listeners under conditions of auditory noise, when their availability can result in improvements in response accuracy equivalent to as much as a 15 dB increase in the auditory signal-to-noise ratio (Grant & Greenberg, 2001; Grant & Seitz, 2000; Sumby & Pollack, 1954). Visual cues can also create some powerful illusions, including the McGurk illusion, where incongruent auditory and visual inputs result in an overall percept derived from but different to the input from each sensory modality (McGurk & MacDonald, 1976). For example a visual /ga/ dubbed over an auditory /ba/ often results in the percept /da/. Other illusions similarly involve visual cues altering the perceived content (Green, Kuhl, Meltzoff & Stevens, 1991) or location (Alais & Burr, 2004) of the auditory signal.

The development of audio-visual speech perception

Work with infants indicates a very early sensitivity to multisensory speech cues. By two months of age infants can match auditory and visual vowels behaviourally (Kuhl & Meltzoff, 1982; Patterson & Werker, 1999). Bristow and colleagues (Bristow, Dehaene-Lambertz, Mattout, Soares, Gilga, Baillet & Mangin, 2008) used an electrophysiological mismatch negativity paradigm to show that visual speech cues habituated 10-week-old infants to auditory tokens of the same phoneme, but not auditory tokens of a different phoneme. Such evidence suggests that infants have a multisensory representation of the phonemes tested, or at least are able to match across senses in the speech domain. By 5 months of age, infants are sensitive to the McGurk illusion, as shown both behaviourally (Burnham & Dodd, 2004; Rosenblum, Schmuckler & Johnson 1997; Patterson & Werker, 1999), and electrophysiologically (Kushnerenko, Teinonen, Volein & Csibra, 2008). Notably though, audio-visual speech perception may not be robust or consistent at this age due to a relative lack of experience (Desjardins & Werker, 2004). Nevertheless, infants pay attention to the mouths of speakers at critical times for language development over the first year (Lewkowicz & Hansen-Tift, 2012), during which time they may even use visual cues to help develop phonemic categories (Teinonen, Aslin, Alku & Csibra, 2008).

By contrast, children do not seem to show sensitivity to, or benefit from, visual cues to the extent that the infant data might predict (e.g. Massaro, Thompson, Barron & Laren, 1986). Typically, children have been shown to be insensitive to the McGurk illusion at age 5, then to show a gradual or stepped developmental progression to the end of primary school or into the teenage years (Hockley & Polka, 1994; McGurk & MacDonald, 1976). Reliable responses to this illusion emerge at around 8 or 9 years (Tremblay, Champoux, Voss, Bacon, Lapore & Theoret, 2007), the same age at which children robustly use visual cues to help overcome noise in the auditory signal (Wightman, Kistler & Brungart, 2006). Ross and colleagues (Ross, Molholm, Blanco, Gomez-Ramirez, Saint-Amour & Foxe, 2011) demonstrated not only the increasing benefit of visual cues over the ages of 5 to 14, but also a change in the profile of how useful visual speech cues were under conditions of different auditory signal-to-noise ratios. Of particular interest in a discussion of developmental trajectories is the finding from an indirect measure of audio-visual speech perception that, while 5-year-olds do not show sensitivity to visual cues, 4-year-olds do (Jerger, Damian, Spence, Tye-Murray & Abdi, 2009); hinting at a U-shaped developmental trajectory in audio-visual speech development.

This developmental pattern of very early sensitivity but late mastery is mirrored in other domains of multisensory development. For example, at 4 months old infants are subject to low-level audio-visual illusions (Kawabe, Shirai, Wada, Miura, Kanazawa & Yamaguci, 2010; Wada, Shirai, Midorikawa, Kanazawa, Dan & Yamaguchi, 2009). However, accuracy in the use of information from multiple senses continues to improve through childhood, and mastering the ability to appropriately weight information from different senses according to their reliability only emerges from around age 8 (Gori, Del Viva, Sandini & Burr, 2008).

Electrophysiological recordings of multisensory speech

The aim of the current work was to understand the development of audio-visual speech perception at the neurophysiological level. Event-related potential (ERP) recordings have repeatedly been used to explore the mechanisms of multisensory processing with adult samples, largely due to the excellent temporal resolution of this technique (Besle, Bertrand & Giard, 2009; Teder-Salejarvi, McDonald, DiRusso & Hillyard, 2002). In this case, we were interested in how visual cues influence, or modulate, auditory processing of speech stimuli. The auditory N1 and P2 ERP components, often referred to together as the *vertex potential*, are highly responsive to auditory speech (e.g. Hoonhorst, Serniclaes, Collet, Colin, Markessis, Radeau & Deltenrea, 2009; Pang & Taylor, 2000). The characteristics of these early-to-mid

latency auditory components, when evoked in response to speech stimuli, are modulated by the presence of visual speech cues in adults (Bernstein, Auer, Wagner & Ponton, 2007; Besle, Fischer, Bidet-Caulet, Lecaignard, Bertrand & Giard, 2008; Besle, Fort, Delpuech & Giard, 2004; Klucharev, Mottonen & Sams, 2003; Pilling, 2009; Stekelenburg & Vroomen, 2007; van Wassenhove, Grant & Poeppel, 2005). Visual cues are shown to both attenuate the amplitude of N1 and P2 as well as, given congruence between auditory and visual inputs, shorten their latency (Pilling, 2009; van Wassenhove *et al*., 2005). While auditory N1 and P2 are most robustly modulated by visual speech, even earlier electrophysiological activity is affected. The auditory P50 is attenuated during intracranial (Besle *et al*., 2008) and sub-dural (Reale, Calvert, Thesen, Jenison, Kawasaki, Oys, Howard & Brugg, 2007) recordings over the lateral superior temporal gyrus; and even auditory brainstem responses and middle latency auditory evoked potentials attenuate in amplitude and reduce in latency when participants are able to see a talking face (Musacchia, Sams, Nicol & Kraus, 2006).

Given multiple replications of the modulation of auditory N1 and P2 by visual speech cues in adults (Bernstein *et al*., 2007; Besle *et al*., 2004, 2008; Klucharev *et al*., 2003; Pilling, 2009; Stekelenburg & Vroomen, 2007; Van Wassenhove et al., 2005), and the correlation of these effects with the perception of multisensory illusions (Van Wassenhove et al., 2005), this can reasonably be taken to represent at least the influence of visual cues on auditory processing, even if not necessarily the integration of information at the single-neuron level. Here we traced these markers of audio-visual speech perception through development. Finding either the modulation of amplitude or latency of the N1/P2 complex over development could help establish the limitations on children's use of multisensory speech cues. Experiment 1 therefore used an adult sample to validate a novel child-friendly paradigm and stimulus set by replicating previous findings of congruence-dependent latency modulation and congruence-independent amplitude modulation of auditory N1 and P2 by visual cues (Van Wassenhove et al., 2005). Four experimental conditions allowed the assessment of the impact of visual speech cues on auditory processing: Auditory-only, Visual-only, congruent Audio-Visual and incongruent audio-visual, referred to as Mismatch. The Mismatch condition was included to assess the effect of audio-visual congruency and to control for a more general effect of attention to the talking face. Experiment 2 used the same paradigm to trace the development of these modulatory effects over mid-to-late childhood, with a sample of children ranging from 6 to 11 years.

Experiment 1

Method

Participants

Participants were 12 native English-speaking adults, who were naive to the experimental hypotheses (mean age = 28.10 years, age range = 20.0–34.0 years). Participants were recruited through the Birkbeck College participant pool and were paid in exchange for taking part. Participants gave their written, informed consent. The experiment was approved by the Birkbeck College Ethics Committee.

Stimuli

When studying auditory ERP components in response to speech stimuli, previous studies have used repetitive consonant-vowel (CV) syllables such as [pa] (e.g. Besle *et al*., 2004) or single vowels (Klucharev *et al*., 2003). Here, the stimulus set was chosen to be as consistent with previous studies as possible while maximizing the likelihood that young children would remain attentive and motivated. The stimuli therefore consisted of a set of monosyllabic, concrete, highly imageable nouns such as 'bell' and 'pen'. The stimuli were recorded by a phonetically trained, female, native English speaker. In total 62 nouns were used, 19 of which were animal names such as 'cat' and 'pig'. The animal names acted as targets during the paradigm and were therefore not included in the ERP analysis. Of the 43 non-target nouns, 31 began with fricatives and three with affricates (of these 18 were bilabial, nine were alveolar and seven were velar), seven stimuli began with liquids and two with a vowel; in total, 29 stimuli began with a voiced phoneme. Sharp acoustic onsets were maintained across the stimulus set as the auditory N1 is sensitive to changes such as rise time (Spreng, 1980). Average age of acquisition of the non-target stimuli was 4.2 years (*SD* = 0.9 years) according to American norms (Kuperman, Stadthagen-Gonzalez & Brysbaert, 2012), and only two of the stimuli ('rose' and 'jam') had an age of acquisition marginally above the age of the youngest participant.

Stimuli were recorded with a digital camera, at 25 frames per second, and a separate audio recording was made simultaneously. Each token was recorded twice and the clearest exemplar was used to create the stimulus set. Auditory tokens were lined up with their corresponding visual tokens by matching the points of auditory onset in the tokens recorded by the external microphone and the video-camera's built-in microphone; auditory recordings were made at a sampling rate of

44.1 kHz. Each token was edited to be 2000 ms long, including an 800 ms period at the start of each clip before auditory onset. There were therefore 800 ms during which visual articulatory cues were available before the onset of auditory information. This allowed for the natural temporal dynamics of audio-visual speech to remain intact while ensuring that each clip began with a neutral face. The length of this period was determined by the clip with the latest auditory onset relative to the onset of natural mouth movements, thus ensuring that no clips were manipulated in order to include this 800 ms visual-only period. The audible portion of each clip lasted on average 437 ms (*SD* = 51 ms).

These tokens were used as the stimulus set for the congruent *Audio-visual* (AV) condition. The stimuli for the three other conditions were then derived from them. A set of *Auditory-only* (AO) and a set of *Visual-only* (VO) stimuli were created by splitting the original tokens into their auditory and visual components. A final set of incongruent audio-visual, *Mismatch* (MM), stimuli were created by mismatching auditory and visual tokens but maintaining the relative timing. For example the auditory token [lake] was dubbed on top of the visual token |rose| 800 ms after its onset. Tokens were paired according to onset phoneme, but such that none resulted in an illusory percept. Animal tokens were kept separate from non-animal tokens when Mismatch stimuli were made, as they were task-relevant.

Procedure

Testing was conducted in an electrically shielded room with dimmed lights. Participants were told that they would either see, hear, or both see and hear a woman saying words and that whenever she said an animal word they should press the mouse button. The button press task was included to help maintain the attention and motivation of the child participants. The role of attention is particularly important here, as the auditory N1 is both amplified and shows more temporal precision with increased selective attention (Martin, Barajas, Fernandez & Torres, 1988; Ritter, Simson & Vaughn, 1988; Thornton, 2008). Stimuli were presented via headphones at approximately 65 dB (SPL), as measured by a sound level meter 2 inches from the centre of the ear pad. Participants were seated in a chair 60 cm from the stimulus presentation screen, and used a chin rest to help keep their heads still and ensure that distance from the screen was kept constant.

Participants completed five blocks of 60 trials. Over the course of five blocks, 75 stimuli of each condition were played, including five animal stimuli per block, resulting in a total of 300 trials per participant. In total, 25 trials were target (animal) trials and were therefore not included in the analysis. The 43 non-target nouns were each repeated either once or twice in each of the four conditions over the course of the experiment. Conditions were randomly presented during each block, although the stimuli presented in each block were the same for each participant. During an audio-visual (AV or MM) or Visual-only (VO) trial a fixation screen appeared for a random period of time between 100 and 400 ms, followed immediately by the video clip, as shown in Figure 1. The fixation variation was intended to minimize expectancy, which has been shown to both attenuate N1 amplitude (Lange, 2009; Viswanathan & Jansen, 2010) and result in slow wave motor anticipatory activity (Teder-Salejarvi *et al.*, 2002). During Auditory-only (AO) trials, the fixation screen remained during the stimulus presentation, after the same jittered period before auditory stimulus onset as for the other conditions. Participants were instructed to remain looking at the centre of the screen at all times, and deviations of gaze were monitored during each session using a video camera. Cartoon eyes on a white background were used as fixation and were located where the bottom of the speaker's nose appeared during video clips. The testing procedure lasted around 45 minutes.

Recording

High density Electrical Geodesics, Inc. (EGI) caps with 128 electrodes joined and aligned according to the international 10–20 system (Jasper, 1958) were used. All bio-electrical signals were recorded using EGI NetAmps (Eugene, OR), with gain set to 10,000 times. The signals were recorded referenced to the vertex (Cz), and were re-referenced to the average during analysis.

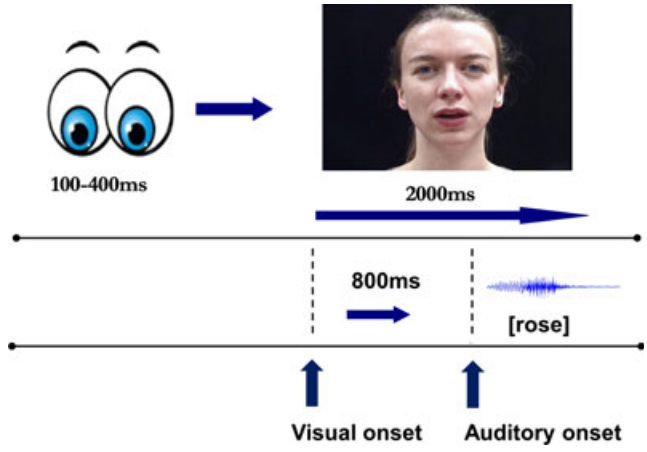

Figure 1 *Example audio-visual trial timeline.*

Data were recorded at 500 Hz and band-pass filtered online between 0.1 and 200 Hz. An oscilloscope and audio monitor were used to measure the accuracy of the relationship between stimulus presentation and electrophysiological recording, and to check the preservation of the relationship between auditory and visual stimuli. No more than 1 ms difference in disparity between audio and visual timing was recorded for any condition.

Analysis and results

Analysis

The region of interest was defined as that which has previously been reported as most appropriate for recording mid-to-late latency auditory ERP components (see e.g. Giard, Perrin, Echallier, Thevenet, Fromenet & Pernier, 1994; Picton, Hillyard, Krausz & Galambos, 1974). The region comprised five channels around, and including, the apex, Cz, which showed the clearest auditory components for these data. The two components analysed at this region of interest were the auditory N1 and auditory P2, with an average of activity taken over the five electrodes. The two ERP measures taken were peak-to-peak amplitude and peak latency for the N1 and P2 components. Windows of analysis were defined as follows: for the P1 (the amplitude of which was used to analyse the N1 component as the N1 and P2 were measured as peak-to-peak values) a window from 40 to 90 ms post stimulus onset was used; for the N1, 80–140 ms; and for the P2, 160–230 ms. The analysis windows were based on a visual inspection of the grand average waveform and checked against data for each individual participant.

Artefact detection was conducted using an automatic algorithm to mark channels as bad if activity exceeded 100 μV at any point; these data were then checked by hand. Trials were rejected if 15 or more channels (12%) were marked as bad. Of those trials included in the analysis, an average of 1.1 channels (0.9%) were marked bad and the data for those channels were interpolated from the remaining channels. Participants were included in the analysis if they contributed at least 30 non-target trials per condition. All adult participants met this condition. The average percentage of trials included per condition was as follows: AO – 79% (*SD* = 16.8), VO – 90% (*SD* = 9.8), AV – 85% (*SD* = 10.4), MM – 83% (*SD* = 13.4).

We directly compared activity in response to the audio-visual conditions with that in response to the AO condition, as only the modulation of auditory responses was of interest for the current purposes. Directly comparing unisensory and multisensory conditions avoids the issue of subtracting activity common to both auditory and visual unimodal responses, which can occur when using the more traditional model of comparing multisensory activity to the sum of the unisensory responses (Stekelenburg & Vroomen, 2007; Teder-Salejarvi *et al*., 2002).

Results

Behavioural results

Accuracy of behavioural responses was converted to d′, with a button press in response to an animal trial counting as a hit and any other button press as a false alarm. Only responses to AO and AV trials are reported here as the main aim of the behavioural task was to maintain attention. VO trials are not reported as the task was not designed to assess lip-reading ability, nor MM trials, due to difficulty in interpretation. The average d′ for AO trials was 3.7 (*SD* = 1.4) and for AV trials was significantly greater ($t(11) = 3.22$, $p = .008$) at 5.7 (*SD* = 2.3). Correlations were run between these behavioural measures and each electrophysiological measure taken, but none reached significance after Bonferroni correction for multiple comparisons.

Electrophysiological results

The adult electrophysiological data followed the same pattern as that seen in previous studies, but with an additional effect of amplitude modulation for the P2 component. A 3 × 2 repeated measures ANOVA was run with three levels of Condition (AO, AV, MM) and two levels of Component (N1 and P2), for amplitude and latency separately. For amplitude, a main effect of Condition was found, $F(2, 22) = 28.43$, $p < .001$, $\eta p^2 = 0.72$, with Bonferroni corrected pairwise comparisons revealing differences ($p < .05$) between each condition, AO > AV > MM. An interaction between Condition and Component also emerged, $F(2, 22) = 9.90$, $p = .001$, $\eta p^2 = 0.47$ with P2, $F(2, 22) = 26.47$, $p < .001$, $\eta p^2 = 0.71$, being more strongly modulated than N1, $F(2, 22) = 16.33$, $p < .001$, $\eta p^2 = 0.60$. Notably, after Bonferroni correction P2 showed significant ($p < .01$) modulation between all levels of Condition, whereas N1 only showed a difference between AO and each audio-visual condition, at $p < .01$ (see Figure 2). For latency, there was a main effect of condition, $F(2, 22) = 4.89$, $p = .017$, $\eta p^2 = 0.31$, driven by the difference ($p < .05$) between the AV condition and the other two conditions, such that AV < AO = MM, given Bonferroni correction for multiple comparisons. Latency modulation was therefore

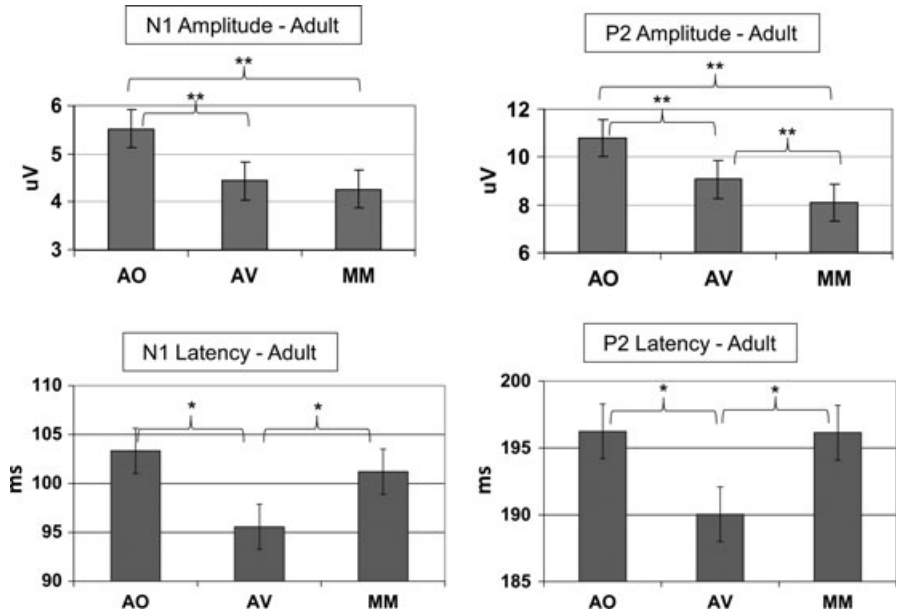

Figure 2 *Peak amplitude and latency of the auditory N1 and P2 components under the Auditory-only (AO), congruent Audio-visual (AV) and Mismatch (MM) conditions, for the adult participants. *p = 0.05; **p = 0.01.*

congruence-dependent. No interaction between condition and component emerged.

Discussion

The aim of experiment 1 was to replicate in adults previous findings of the modulation of auditory ERP components by visual speech cues using a child-friendly paradigm and stimulus set.

Adult use of visual cues

Compared to auditory-only speech stimuli, audio-visual stimuli resulted in congruence-independent attenuation of N1 and P2 component amplitude and congruence-dependent shortening of component latency. The modulation of auditory ERP components therefore replicated previous findings (Pilling, 2009; Van Wassenhove et al., 2005). This data set validated the use of the child-friendly paradigm on adults for subsequent use with a developmental sample.

van Wassenhove and colleagues (2005) proposed that the shortening of component latency in the presence of visual speech cues represents the use of visual cues to predict the content of the upcoming auditory signal; a proposal known as the 'predictive coding hypothesis'. This is possible in natural speech as the onset of visual cues occurs between 100 and 300 ms before their auditory counterparts (Chandrasekaran *et al.*, 2009). van Wassenhove and colleagues found particularly strong support for this notion as latency shortening was not only sensitive to congruency but further to the degree of ambiguity of the onset phoneme. Greater latency modulation was recorded given the syllable [pa] over [ta] and given [ta] over [ka]. In this study |pa| was the least ambiguous viseme (the visual correlate of an auditory phoneme), and as such was suggested to make a stronger prediction and result in faster processing of the more expected auditory signal. Ease of processing has previously been associated with the shortening of auditory N1 latency (Callaway & Halliday, 1982). Although stimuli in the current study could not be analysed by onset phoneme, the congruence-dependent shortening of latency further supports the predictive coding hypothesis.

We additionally replicated findings of amplitude modulation regardless of congruency between the auditory and visual inputs, driven predominantly by the P2 component (Pilling, 2009; Van Wassenhove et al., 2005). Two hypotheses have been put forward in the literature to explain congruence-independent effects of one sensory modality on another. van Wassenhove and colleagues suggested that a reduction of amplitude results from visual speech cues driving more efficient auditory processing. The authors proposed that redundant information, carried in both senses, need not be fully processed by the auditory system, resulting in more efficient processing of information available through the auditory channel. In the case of visual speech cues, this may entail a reduction in processing of information from the second

and third formants, which carry information about place of articulation.

An alternative explanation, known as the 'deactivation hypothesis' (Bushara, Hanawaka, Immisch, Toma, Kansaku & Hallett, 2003; Wright, Pelphrey, Allison, McKeown & McCarthy, 2003), asserts that different parts of the multisensory processing stream are in competition, such that stimuli from different senses showing temporal and spatial synchrony produce super-additive activity in some areas, but suppression of activity in others. Under this view, when multisensory stimuli are available, regions that process more than one sense dominate over unisensory areas. So, for example, responses in auditory cortex are reduced in the presence of visual information about the same object or event, as multisensory processing regions compete and dominate. Experimental evidence from fMRI studies supports the theoretical notion of competition between unisensory and multisensory areas (Bushara *et al*., 2003).

However, in the current data set the attenuation of P2 amplitude was greater for the audio-visual Mismatch condition than for the congruent Audio-visual condition. Given that an incongruent visual cue does not provide more reliable information regarding place of articulation, nor does it result in the perception of a multisensory event, these data are difficult to reconcile with either of the above hypotheses. A possible explanation lies in the nature of the stimuli used here. In the current study, the Mismatch stimuli consisted of entirely unrelated words presented in each sensory modality, for example, auditory [lake] paired with visual |rose| This is in contrast to previous studies which have used McGurk stimuli (Pilling, 2009; Van Wassenhove et al., 2005), that is, incongruent CV syllables which can form coherent percepts despite their physical mismatch.

The current data therefore support an alternative hypothesis that amplitude attenuation reflects competition between sensory inputs, with competition being greater when auditory and visual systems are processing incompatible, and irreconcilable, stimuli. That this effect is restricted to the P2 component is compatible with evidence that it originates in posterior superior temporal cortex (Liebenthal, Desai, Ellinson, Ramachandran, Desai & Binder, 2010). The posterior superior temporal cortex is composed of the posterior superior temporal gyrus (pSTG) and sulcus (pSTS) and forms part of a network of regions implicated in audio-visual speech processing. This network also includes primary sensory cortices, frontal and pre-motor regions and the supramarginal gyrus (see Campbell, 2008, for a review). The pSTS is the most reliably activated region in fMRI studies in response to audio-visual over auditory speech, and lip-reading (Calvert, Bullmore, Brammer, Campbell, Woodruff, McGuire, Williams, Iversen & David, 1997; Calvert, Campbell & Brammer, 2000; Callan, Jones, Munhall, Kroos, Callan & Vatikiotis-Bateson, 2004; Capek, Bavelier, Corina, Newman, Jezzard & Neville, 2004; Hall, Fussell & Summerfield, 2005; Skipper, Nusbaum & Small, 2005). Furthermore, pSTS is associated with learning inter-sensory pairings (Tanabe, Honda & Sadato, 2005), with auditory expertise (Leech, Holt, Devlin & Dick, 2009) and shows sensitivity to congruency in ongoing audio-visual speech (Calvert *et al*., 2000). In a systematic analysis of the role of pSTS in audio-visual processing, Hocking and Price (2008) suggest that this region is involved in conceptual matching regardless of input modality.

Given that cortical regions involved in the generation of the auditory P2 component are sensitive to matching auditory and visual stimuli, the attenuation of P2 may reflect competition between neurons in a multisensory population responsive to different modalities, with competition increasing given irreconcilable incongruence. A possible next step in the examination of this hypothesis is to compare reconcilable (i.e. McGurk) and irreconcilable incongruent audio-visual speech stimuli within the same paradigm.

Experiment 2

Experiment 2 traced the developmental trajectory of auditory ERP modulation by visual speech cues from age 6 to 12, over which period children establish a reliable use of visual cues to aid speech perception as shown using behavioural measures (e.g. Wightman *et al*., 2006). We sought to determine whether modulation of ERPs due to multisensory processing could be observed at an earlier age than has been measured behaviourally.

Method

Participants

Thirty-eight typically developing children participated (mean age = 8.9 years, *SD* = 21 months, age range = 6.0–11.10 years, with between five and seven children in each year group). Children were recruited by placing advertisements in the local press, and were rewarded for their participation with small toys. Parents gave written, informed consent for their children. The experiment was approved by the Birkbeck College Ethics Committee. One child was excluded from the analysis as a result of excessive noise in the data.

Recording and procedure

The experimental procedure for children was almost identical to that used in Experiment 1 for adult participants. The procedure lasted slightly longer for children, around 60 minutes, as more time was spent practising sitting still. Blinking was not mentioned as it was judged that this would be hard for young children to control and would only serve to draw attention to the act. Paediatric EGI electroencephalographic nets with 128 electrodes were used for all child participants.

Analysis

The same region of interest and the same epoch windows were used for the child sample based on grand average data for each age group and checked against data for each individual participant. After artefact rejection, slightly more data were discarded as noisy than for the adult sample. For child participants, an average of 3.6 channels (2.8%) were marked bad on accepted trials. As per the adults, participants were included in the analysis if they contributed at least 30 non-target trials per condition; one child was excluded from analysis on these grounds. The average percentage of trials included for the child sample was: AO – 57% (*SD* = 14.4), VO – 73% (*SD* = 12.6), AV – 68% (*SD* = 14.7), MM – 67% (*SD* = 13.9).

Results

Behavioural results

The average d′ for the child sample was 2.5 (*SD* = 1.9) for AO and 2.7 (*SD* = 1.9) for AV trials. d′ was consistently good, with each age group scoring significantly above zero on each measure at $p < .05$, indicating satisfactory attention across all ages. Behavioural performance improved over developmental time, with Age predicting performance on both AO ($R^2 = 0.19$, $F(1, 35) = 8.26$, $p = .007$) and AV trials ($R^2 = 0.15$, $F(1, 35) = 6.11$, $p = .018$). Unlike the adult sample in Experiment 1, on this simple detection task the child sample showed no behavioural benefit of AV trials over AO trials. Correlations between behavioural d′ and brain responses were calculated for the child sample, but again no correlations survived Bonferroni correction for multiple comparisons.

Electrophysiological results

Figure 3 shows the grand average waveforms for the 6- and 7-year-olds, the 8- and 9-year-olds, the 10- and 11-year-olds as well as the adults from Experiment 1,

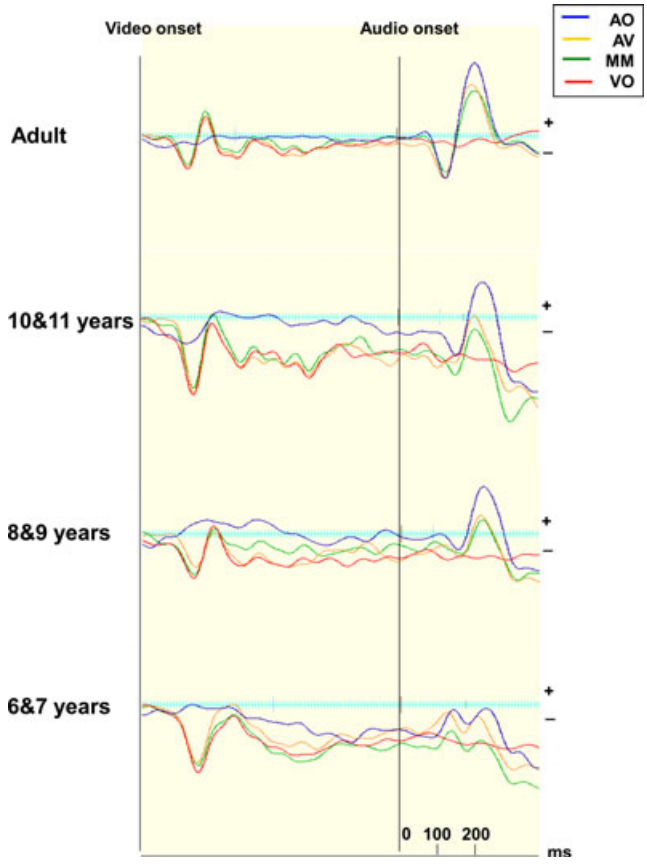

Figure 3 *Grand average waveforms for each condition, Auditory-only (AO), Audio-visual (AV), Mismatch (MM) and Visual-only (VO) at the region of interest. Waveforms are shown divided by age group. The onset points of the visual and auditory stimuli are shown.*

with the amplitude and latency values for the auditory N1 and P2 components shown in Table 1. These categorical age groupings are used here to illustrate developmental change but in further analyses age is treated as a continuous variable. To assess change over time, the developmental data were entered into a repeated measures ANCOVA with Condition (AO, AV, MM), and Component (N1, P2) as the within subjects factors, and Age (in months) added as a covariate. Main effects of Condition were analysed separately in an ANOVA (see Thomas, Annaz, Ansari, Serif, Jarrold & Karmiloff-Smith, 2009).

A main effect of Condition emerged, $F(2, 72) = 10.16$, $p < .001$, $\eta p^2 = 0.22$, with Bonferroni corrected pairwise comparisons revealing differences ($p < .01$) between AO and each multisensory condition, AO > AV = MM. An interaction between Condition and Component emerged, $F(2, 72) = 9.59$, $p < .001$, $\eta p^2 = 0.21$, with the P2 component being effected by Condition, $F(2, 72) = 17.12$,

Table 1 *Means (and standard deviations) for auditory N1 and P2 amplitude (peak to peak) and peak latency, for each age group. Latency values are not given for the Visual-only condition, as amplitude values show latent activity within the window of analysis rather than components*

		Auditory- only	Visual- only	Audio-visual	Mismatch
N1 amplitude (μV)	6&7	2.5 (1.6)	2.3 (1.4)	3.3 (2.2)	2.3 (1.6)
	8&9	3.5 (3.7)	1.6 (0.9)	2.9 (2.1)	3.1 (3.6)
	10&11	3.7 (1.3)	1.6 (0.9)	2.3 (1.1)	3.0 (1.4)
	Adult	5.5 (1.4)	1.5 (0.4)	4.4 (1.2)	4.3 (1.4)
N1 latency (ms)	6&7	114.8 (14.7)	–	114.1 (10.5)	117.0 (11.1)
	8&9	105.5 (10.7)	–	105.8 (12.4)	110.4 (14.0)
	10&11	109.1 (12.5)	–	102.0 (12.7)	105.0 (11.4)
	Adult	103.3 (11.1)	–	95.6 (11.0)	101.2 (7.0)
P2 amplitude (μV)	6&7	3.4 (2.3)	1.5 (1.1)	2.9 (2.3)	2.2 (2.1)
	8&9	6.6 (3.7)	1.8 (1.2)	4.7 (2.8)	4.2 (4.0)
	10&11	6.5 (2.7)	1.5 (1.4)	4.2 (2.9)	3.9 (2.2)
	Adult	10.8 (2.9)	1.5 (0.6)	9.1 (2.0)	8.1 (2.1)
P2 latency (ms)	6&7	195.2 (15.5)	–	182.6 (17.1)	187.0 (20.2)
	8&9	188.9 (11.6)	–	182.0 (13.9)	183.5 (9.6)
	10&11	195.1 (17.6)	–	183.3 (19.8)	180.4 (11.7)
	Adult	196.2 (8.7)	–	190.0 (9.5)	196.1 (12.4)

$p < .001$, $\eta p^2 = 0.32$, but not the N1 ($p = .420$). Again this P2 effect was driven by the difference ($p < .001$) between AO and each multisensory condition (AO > AV = MM), as shown by Bonferroni corrected pairwise comparisons.

There was no main effect of Age, but there was a significant interaction between Age and both Component, $F(1, 35) = 9.52$, $p = .004$, $\eta p^2 = 0.21$, and Condition, $F(2, 70) = 4.05$, $p = .022$, $\eta p^2 = 0.10$. The first of these interactions was driven by the P2 component showing a main effect of Age, $F(1, 35) = 5.31$, $p = .027$, $\eta p^2 = 0.13$, whereas the N1 component did not ($p = .991$). The Age by Condition interaction was driven by the AO condition showing a main effect of Age, $F(1, 35) = 4.14$, $p = .050$, $\eta p^2 = 0.11$, but not the AV ($p = .97$) or the MM ($p = .198$) conditions. So, the main effect of Condition revealed by the ANOVA seems to have been driven predominantly by the older children, and as a result of the AO response getting larger over development (as illustrated in Figure 4).

To further assess the changing relationship between Conditions over Age, a linear regression was run with Age as a predictor of the difference between AO and each audio-visual condition for N1 and P2. Age was found to significantly predict the difference between the AO and AV conditions for N1 amplitude, $R^2 = 0.13$, $F(1, 35) = 5.38$, $p = .026$, $\beta = 0.365$, and P2 amplitude, $R^2 = 0.13$, $F(1, 35) = 5.077$, $p = .031$, $\beta = 0.356$. The age at which the difference between conditions became significant was determined using the 95% confidence intervals around the regression lines (see Figure 5). The lower boundary crossed zero at 122 months (10.1 years) for N1 amplitude, and at 89 months (7.4 years) for P2 amplitude. The increasing difference between conditions was approximately equivalent for each component. However, Figure 4 suggests that for the N1 component, the change in difference results predominantly from a decrease in Audio-visual response amplitude, while for P2 the change was predominantly driven by an increase in Auditory-only amplitude. Age did not predict the difference between the AO and MM conditions for either the N1 ($p = .846$) or P2 ($p = .087$) components.

For latency, the ANOVA revealed a main effect of Condition, $F(2, 72) = 5.14$, $p = .008$, $\eta p^2 = 0.13$, driven by the difference ($p < .05$) between the AO and each audio-visual condition, AO > AV = MM. An interaction also emerged between Condition and Component, $F(2, 72) = 5.52$, $p = .006$, $\eta p^2 = 0.13$. The P2 component was significantly influenced by Condition, $F(2, 72) = 7.30$, $p = .001$, $\eta p^2 = 0.17$, driven by the Bonferroni corrected difference ($p < .05$) between AO and both audio-visual conditions; the N1 component was not influenced by Condition ($p = .128$).

The ANCOVA for latency revealed a main effect of Age, $F(1, 35) = 4.56$, $p = .040$, $\eta p^2 = 0.12$, but no interaction between Age and Condition (see Figure 4.). So, the latency of these auditory components was seen to shorten over development, but the effect of Condition did not change over this age range.

All analyses were re-run comparing responses to the multisensory conditions with responses to the sum of the unisensory conditions. This is a more traditional approach adopted in multisensory processing studies (see Calvert, 2001). The results of this analysis showed the same pattern but with larger sub-additive effects, that is, the effect of Condition was exaggerated for all comparisons and was therefore less conservative.

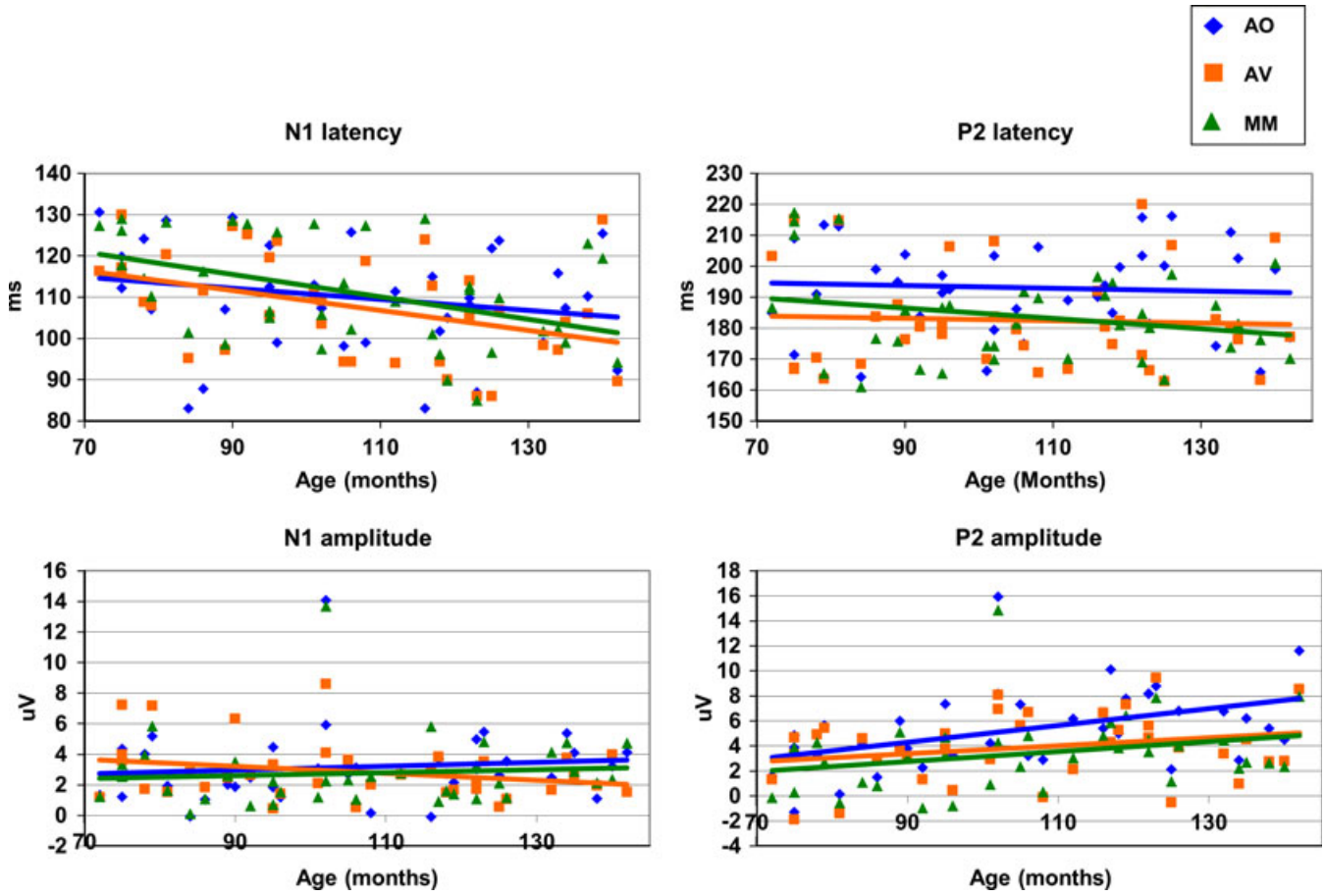

Figure 4 *Developmental trajectories for the Auditory-only (AO), Audio-visual (AV) and Mismatch (MM) conditions for auditory N1 and P2 peak to peak amplitude and peak latency.*

Discussion

The influence of visual cues over mid-to-late childhood

With regard to amplitude, as a group, the children responded similarly to the adults, in that the P2 component was attenuated given congruent and incongruent visual cues compared to the Auditory-only condition. Over developmental time the P2 component increased in amplitude, with this effect being driven by an increase in response to the Auditory-only condition. Age predicted the difference between the Auditory-only and Audio-Visual (congruent) conditions for both components, with this effect on P2 predominantly resulting from an increased response to the Auditory-only stimuli, while for the N1 component a slight decrease in amplitude in response to the Audio-visual stimuli seems to be responsible. The difference between conditions became significant from 10.1 years for the N1 component, and at 7.4 years for the P2 component. The period between these two ages matches that seen in behavioural studies when visual speech cues come to reliably influence auditory perception both in terms of the McGurk illusion and audio-visual advantage during speech-in-noise (e.g. Tremblay *et al.*, 2007; Wightman *et al.*, 2006). These results suggest that the modulation of different auditory components represents separate processes in the integration and/or use of visual speech cues, and that this developmental process may be traced at the behavioural level. What is not clear is exactly what the information processing correlates of N1 and P2 attenuation might be.

If amplitude modulation does represent competition between inputs from different sensory modalities, as suggested above, then the developmental data imply that this response only emerges over mid-to-late childhood, but is not fully mature by age 12 as the additional amplitude attenuation seen in adults to incongruent audio-visual stimuli was not seen for the oldest children in this sample. This protracted period of maturation maps onto imaging data showing regions in superior temporal cortex, which contribute to P2 generation in children as they do in adults (Ponton, Eggermont, Khosla, Kwong & Don, 2002), do

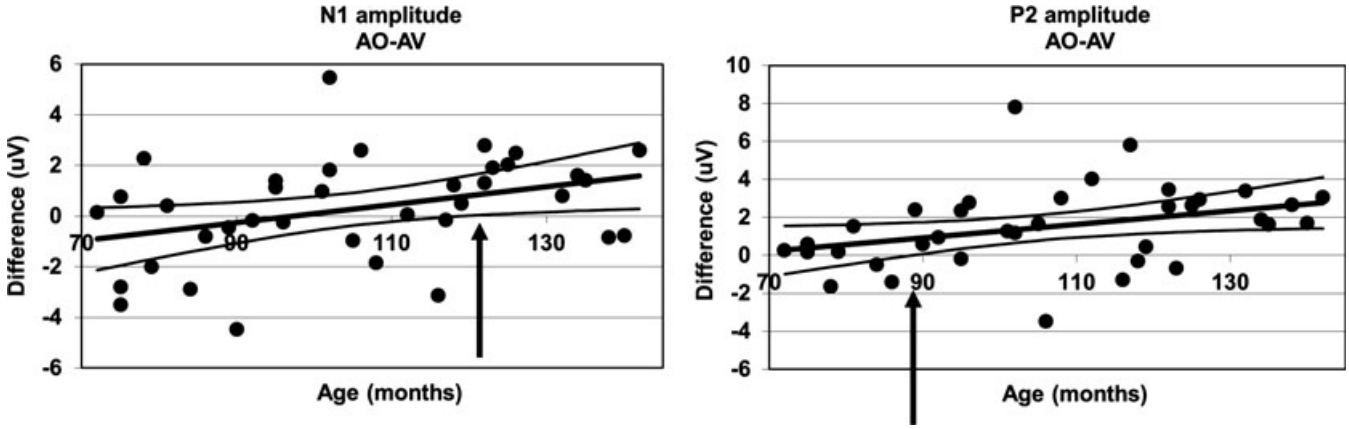

Figure 5 *Regression model with age predicting the difference between the AO and audio-visual conditions for auditory N1 and P2 amplitude. The arrows show the points at which the lower 95% confidence interval crosses 0 (122 and 89 months, respectively).*

not mature until the teenage years (Gotgay *et al.*, 2004; see Lenroot & Giedd, 2006). Recent functional imaging data mirror this late development and support the role of STS in children's audio-visual speech perception (Nath, Fava & Beauchamp, 2011). Dick and colleagues (Dick, Solodkin & Small, 2010) measured brain activity in response to auditory and audio-visual speech in adults and 8- to 11-year-old children, and found that while the same areas were involved in perception for both adults and children, the relationships between those areas differed. For example, the functional connectivity between pSTS and frontal pre-motor regions was stronger for adults given audio-visual over auditory-only speech, but weaker for children.

With regard to latency, a different pattern emerged for the children, as a group, compared to the adult sample in Experiment 1. For the children, only the P2 component exhibited latency modulation in response to visual speech cues, and latency shortening was observed regardless of congruency between auditory and visual cues. Interpretations of previous adult data (Pilling, 2009; Van Wassenhove et al., 2005) have rested on the effect of congruence-dependency, with congruent visual cues suggested to allow a prediction of the upcoming auditory signal, such that the degree of latency shortening reflects the difference between expected and perceived events. The current developmental data are not sensitive to congruency, and therefore cannot be interpreted entirely with recourse to the prediction of signal content. The present and previous adult data may therefore not tell the whole story regarding latency modulation. One possibility is that visual cues are involved in predicting not just *what* is about to be presented, but also *when* it is to be presented. Certainly, using non-speech stimuli, the auditory N1 and P2 components have been shown to be sensitive to both the content and timing of stimulus presentation (Viswanathan & Jansen, 2010). In this case, children of the age range tested here may use visual cues to predict the timing but not the content of the upcoming auditory signal.

The idea that visual speech cues may allow a prediction of when important information in the auditory stream will be presented has been proposed before under the 'peak listening' hypothesis (Kim & Davis, 2004). This theory states that visual speech cues predict when in the auditory signal energetic peaks will occur, which are particularly beneficial when processing speech in noise. If the shortening of latency does represent two predictive measures, then future work should reveal that latency shortening is sensitive to manipulations of both predictability of content and timing of the auditory signal relative to visual cues. Age did not interact with Condition with respect to latency modulation, so no change in the ability to predict the upcoming auditory stimulus emerged over this developmental window. The influence of visual speech cues on the latency of auditory components from age 6 may therefore represent an aspect of audio-visual speech perception that is continuous from infancy despite the U-shaped behavioural trajectory outlined in the introduction. However, the change in congruency dependence must occur after the age of 12, possibly revealing a much later sensitivity to upcoming auditory content.

Over developmental time, a main effect of age on component latency was revealed, indicating that children process these stimuli more rapidly as they get older. Auditory ERP responses are known to show a gradual course of developmental change and maturation over childhood and adolescence (Bishop, Hardiman, Uwer & von Suchodeltz, 2007; Lippe, Kovacevic & McIntosh, 2009). It is hard to tease apart the extent to which these changes result from the slow physiological maturation of the auditory cortex (Moore, 2002), or changes in

cognitive processes functionally underlying the activity or, more likely, a complex interaction between the two.

General discussion

The aim of this study was to chart the trajectory of the modulation of auditory ERP components by visual speech cues over developmental time. We first validated a new child-friendly paradigm using adult participants, which replicated previous findings of congruence-dependent shortening of ERP component latency and congruence-independent attenuation of component amplitude. A greater attenuation of amplitude emerged given mismatched visual speech cues, suggesting that attenuation may represent competition between inputs from different sensory modalities. This competition may be important for the process of evaluating the nature of multisensory stimuli in order to determine whether information across modalities refers to the same object or event. We have shown that the modulation of auditory ERP components by visual speech cues gradually emerges over developmental time, maturing at around the age when behavioural studies have revealed a use of visual cues in speech perception tasks. Notably though, the additional sensitivity to incongruent visual cues seen in adults was not evident in this developmental sample.

Regarding latency shortening, our adult results replicated previous findings, supporting the notion that latency modulation represents the process of predicting the content of the upcoming auditory signal, the predictive coding hypothesis. However, data from our child sample showed latency shortening for the P2 component regardless of the congruence between auditory and visual signals. We have therefore suggested that latency shortening may represent two predictive processes, relating to both the content and timing of the upcoming auditory signal, but that children within the age range tested here are not yet able to make content predictions.

Overall, these data support and extend previous studies pointing to the influence of visual cues on processing auditory speech. We have supported the notion that amplitude and latency modulation represent different aspects of audio-visual signal processing, but reinterpreted those data in the light of our new paradigm, and the developmental results. Furthermore, we have presented new data revealing that these responses gradually emerge over childhood.

Study limitations and outstanding questions

This study was successful in its aim to develop a child-friendly ERP paradigm for the study of audio-visual speech, but was limited in a number of respects. The age range tested here, although relatively wide, was not sufficient to fully trace the development of the electrophysiological markers of audio-visual speech perception into adulthood. Another limitation, in terms of being able to draw firm conclusions, was that the audio-visual Mismatch stimuli used here were all irreconcilably incongruent. While this led to an interesting finding when compared to previous studies with adults, it might also have changed the strategy of participants. As matched and mismatched multisensory stimuli were randomly intermixed within each block, participants may have adopted more of a 'wait and see' strategy than they would under more naturalistic settings. One way for future studies to address whether this factor had a significant impact on the results would be to separate conditions by block.

Finally it should be noted that all the stimuli here were presented under conditions of no notable auditory noise. This factor may turn out to substantially impact on electrophysiological data given that dynamic functional changes in connectivity have been recorded between unisensory cortices and the STS as a function of noise (Nath & Beauchamp, 2011). This modulation is thought to reflect changes in the weighting of information from each sensory modality, and should be considered in future electrophysiological investigations.

One question that has emerged from the current work is exactly what the development of electrophysiological responses represents at the level of information processing. The data on amplitude modulation presented here fit well with the behavioural data examining the gross benefit of visual cues to children. However, the modulation of component latency was evident at younger ages, and certainly the use of visual cues in infancy suggests that the process is one of continuous change rather than simply 'coming online' later in childhood. This developmental profile may represent changes in how visual speech cues are utilized in childhood with increasing experience and cortical maturation. For example, Fort and colleagues (Fort, Spinelli, Savariaux & Kandel, 2012) found that during a vowel monitoring task both adults and 5- to 10-year-old children, as a group, benefited from the availability of visual speech cues, but only adults showed an additional benefit of lexicality. These authors suggest that where adults use visual cues to help retrieve lexical information, children use the same cues to process the phonetic aspects of speech. Over developmental time, then, children may first use visual speech cues to aid phonetic processing, and later to aid comprehension.

This critical issue of the changing relationship between brain and behaviour over development needs to be addressed with further electrophysiological exploration

in conjunction with more sensitive behavioural methods aimed at elucidating the different potential uses of visual speech cues. The exploration of audio-visual speech over childhood is important not just for typically developing children learning about the world in auditory noise, but also critically for those children growing up with developmental language disorders, for whom multisensory cues may contain valuable information to assist language development.

Acknowledgements

This work was supported by an ESRC studentship awarded to Victoria C.P. Knowland, and an ESRC grant, ERS-062-23-2721, awarded to Michael S.C. Thomas. The authors would like to thank Torsten Baldeweg for his helpful comments.

References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near optimal bimodal integration. *Current Biology*, **14** (3), 257–262.

Bernstein, L.E., Auer, E.T., Wagner, M., & Ponton, C.W. (2007). Spatiotemporal dynamics of audio-visual speech processing. *NeuroImage*, **39**, 423–435.

Besle, J., Bertrand, O., & Giard, M.H. (2009). Electrophysiological (EEG, sEEG, MEG) evidence for multiple audio-visual interactions in the human auditory cortex. *Hearing Research*, **258**, 143–151.

Besle, J., Fischer, C., Bidet-Caulet, A., Lecaignard, F., Bertrand, O., & Giard, M.-H. (2008). Visual activation and audio-visual interactions in the auditory cortex during speech perception: intercranial recording in humans. *Journal of Neuroscience*, **28** (52), 14301–14310.

Besle, J., Fort, A., Delpuech, C., & Giard, M.-H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience*, **20**, 2225–2234.

Bishop, D.V.M., Hardiman, M., Uwer, R., & von Suchodeltz, W. (2007). Maturation of the long-latency auditory ERP: step function changes at start and end of adolescence. *Developmental Science*, **10** (5), 565–575.

Bristow, D., Dehaene-Lambertz, G., Mattout, J., Soares, C., Gilga, T., Baillet, S., & Mangin, F. (2008). Hearing faces: how the infant brain matches the face it sees with the speech it hears. *Journal of Cognitive Neuroscience*, **21** (5), 905–921.

Burnham, D., & Dodd, B. (2004). Auditory-visual speech integration by pre-linguistic infants: perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology*, **44** (4), 204–220.

Bushara, K.O., Hanawaka, T., Immisch, I., Toma, K., Kansaku, K., & Hallett, M. (2003). Neural correlates of cross-modal binding. *Nature Neuroscience*, **6** (2), 190–195.

Callan, D.E., Jones, J.A., Munhall, K., Kroos, C., Callan, A.M., & Vatikiotis-Bateson, E. (2004). Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. *Journal of Cognitive Neuroscience*, **16**, 805–816.

Callaway, E., & Halliday, R. (1982). The effect of attentional effort on visual evoked potential N1 latency. *Psychiatry Research*, **7**, 299–308.

Calvert, G. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cerebral Cortex*, **11** (12), 1110–1123.

Calvert, G.A., Bullmore, E., Brammer, M.J., Campbell, R., Woodruff, P., McGuire, P., Williams, S., Iversen, S.D., & David, A.S. (1997). Activation of auditory cortex during silent speechreading. *Science*, **276**, 593–596.

Calvert, G.A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in human heteromodal cortex. *Current Biology*, **10**, 649–657.

Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. *Philosophical Transactions of the Royal Society*, B, **363**, 1001–1010. doi: 10.1098/rstb.2007.2155.

Capek, C.M., Bavelier, D., Corina, D., Newman, A.J., Jezzard, P., & Neville, H.J. (2004). The cortical organization of audio-visual sentence comprehension: an fMRI study at 4 Tesla. *Cognitive Brain Research*, **20**, 111–119.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLOS Computational Biology*, **5** (7), e1000436.

Desjardins, R., & Werker, J.F. (2004). Is the integration of heard and seen speech mandatory for infants? *Developmental Psychobiology*, **45** (4), 187–203.

Dick, A.S., Solodkin, A., & Small, S. (2010). Neural development of networks for audiovisual speech comprehension. *Brain and Language*, **114** (2), 101–114.

Fort, M., Spinelli, E., Savariaux, C., & Kandel, S. (2012). Audiovisual vowel monitoring and the word superiority effect in children. *International Journal of Behavioral Development*, **36** (6), 457–467.

Giard, M.-H., Perrin, F., Echallier, J.F., Thevenet, M., Fromenet, J.C., & Pernier, J. (1994). Dissociation of temporal and frontal components in the human auditory N1 wave: a scalp current density and dipole model analysis. *Electroencephalography & Clinical Neurophysiology*, **92**, 238–252.

Gori, M., Del Viva, M., Sandini, G., & Burr, D.C. (2008). Young children do not integrate visual and haptic form information. *Current Biology*, **18** (9), 694–698.

Gotgay, N., Giedd, J., Lusk, L., Hayashi, K.M., Greenstein, D., Vaituzis, A.C., Nugent, T.F., III, Herman, D.H., Clasen, L.S., Toga, A.W., Rapopport, J.L., & Thompson, P.M. (2004). Dynamic mapping of human cortical development during childhood through early adulthood. *Proceedings of the National Academy of Sciences, USA*, **101** (21), 1874–1879.

Grant, K.W., & Greenberg, S. (2001). Speech intelligibility derived from asynchronous processing of auditory-visual information. *Proceedings of the Workshop on Audio-Visual Speech Processing* (AVSP-2001).

Grant, K.W., & Seitz, P.F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America*, **108** (3), 1197–1208.

Green, K.P., Kuhl, P.K., Meltzoff, A.N., & Stevens, E.B. (1991). Integrating speech information across talkers, gender, and sensory modality: female faces and male voices in the McGurk effect. *Perception & Psychophysics*, **50**, 524–536.

Hall, D.A., Fussell, C., & Summerfield, A.Q. (2005). Reading fluent speech from talking faces: typical brain networks and individual differences. *Journal of Cognitive Neuroscience*, **17**, 939–953.

Hoonhorst, I., Serniclaes, W., Collet, G., Colin, C., Markessis, E., Radeau, M., & Deltenrea, P. (2009). N1b and Na subcomponents of the N100 long latency auditory evoked--potential: *N*europhysiological correlates of voicing in French-speaking subjects. *Clinical Neurophysiology*, **120** (5), 897–903.

Hocking, J., & Price, C.J. (2008). The role of the posterior temporal sulcus in audiovisual processing. *Cerebral Cortex*, **18** (10), 2439–2449.

Hockley, N.S., & Polka, L. (1994). A developmental study of audiovisual speech perception using the McGurk paradigm. *Journal of the Acoustical Society of America*, **96** (5), 3309–3309.

Jasper, H.H. (1958). The ten–twenty electrode system of the International Federation. *Electroencephalography and Clinical Neurophysiology*, **10**, 371–375.

Jerger, S., Damian, M.F., Spence, M.J., Tye-Murray, N., & Abdi, H. (2009). Developmental shifts in children's sensitivity to visual speech: a new multisensory picture word task. *Journal of Experimental Child Psychology*, **102**, 40–59.

Kawabe, T., Shirai, N., Wada, Y., Miura, K., Kanazawa, S., & Yamaguci, M.K. (2010). The audiovisual tau effect in infancy. *PLoS ONE*, **5** (3): e9503. doi: 10.1371/journal.pone.0009503

Kim, J., & Davis, C. (2004). Investigating the audio-visual speech detection advantage. *Speech Communication*, **44**, 19–30.

Klucharev, K., Mottonen, R., & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audio-visual speech perception. *Cognitive Brain Research*, **18**, 65–75.

Kuhl, P., & Meltzoff, A. (1982). The bimodal perception of speech in infancy. *Science*, **218**, 1138–1141.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age of acquisition ratings for 30,000 English words. *Behavioural Research Methods*, **44** (4), 978–990.

Kushnerenko, E., Teinonen, T., Volien, A., & Csibra, G. (2008). Electrophysiological evidence of illusory audiovisual speech percept in human infants. *Proceedings of the National Academy of Sciences, USA*, **105** (32), 11442–11445.

Lange, K. (2009). Brain correlates of early auditory processing are attenuated by expectations for time and pitch. *Brain and Cognition*, **69** (1), 127–137.

Leech, R., Holt, L.L., Devlin, J.T., & Dick, F. (2009). Expertise with artificial non-speech sounds recruits speech-sensitive cortical regions. *Journal of Neuroscience*, **29** (16), 5234–5239.

Lenroot, R.K., & Giedd, J.N. (2006). Brain development in children and adolescents: insights from anatomical magnetic resonance imaging. *Neuroscience and Biobehavioral Reviews*, **30**, 718–729.

Lewkowicz, D.J., & Hansen-Tift, A.M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences, USA*, **109** (5), 1431–1436.

Liebenthal, E., Desai, R., Ellinson, M.M., Ramachandran, B., Desai, A., & Binder, J.R. (2010). Specialisation along the left superior temporal sulcus for auditory categorisation. *Cerebral Cortex*, **20** (12), 2958–2970.

Lippe, S., Kovacevic, N., & McIntosh, A.R. (2009). Differential maturation of brain signal complexity in the human auditory and visual system. *Frontiers in Human Neuroscience*, **3**, 48.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746–748.

Martin, L., Barajas, J.J., Fernandez, R., & Torres, E. (1988). Auditory event-related potentials in well-characterized groups of children. *Electroencephalography and Clinical Neurophysiology: Evoked Potentials*, **71**, 375–381.

Massaro, D., Thompson, L., Barron, B., & Laren, E. (1986). Developmental changes in visual and auditory contributions to speech perception. *Journal of Experimental Child Psychology*, **41**, 93–113.

Moore, J.K. (2002). Maturation of human auditory cortex: implications for speech perception. *Annals of Otology, Rhinology and Laryngology*, **111**, 7–10.

Musacchia, G., Sams, M., Nicol, T., & Kraus, N. (2006). Seeing speech affects acoustic information processing in the human brainstem. *Experimental Brain Research*, **168** (1–2), 1–10.

Nath, A.R., & Beauchamp, M.S. (2011). Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *Journal of Neuroscience*, **31** (5), 1704–1714.

Nath, A.R., Fava, E.E., & Beauchamp, M.S. (2011). Neural correlates of interindividual differences in children's audiovisual speech perception. *Journal of Neuroscience*, **31** (39), 13963–13971.

Pang, E.W., & Taylor, M.J. (2000). Tracking the development of the N1 from age 3 to adulthood: an examination of speech and non-speech stimuli. *Clinical Neurophysiology*, **111** (3), 388–397.

Patterson, M., & Werker, J. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behavior and Development*, **22**, 237–247.

Picton, T.W., Hillyard, S.A., Krausz, H.I., & Galambos, R. (1974). Human auditory evoked potentials. *Electroencephalography and Clinical Neurophysiology*, **36**, 179–190.

Pilling, M. (2009). Auditory event-related potentials (ERPs) in audio-visual speech perception. *Journal of Speech, Language and Hearing Research*, **52**, 1073–1081.

Ponton, C., Eggermont, J.J., Khosla, D., Kwong, B., & Don, M. (2002). Maturation of human central auditory system activity: separating auditory evoked potentials by dipole source modeling. *Clinical Neurophysiology*, **113**, 407–420.

Reale, R.A., Calvert, G.A., Thesen, T., Jenison, R.L., Kawasaki, H., Oys, H., Howard, M.A., & Brugg, J.F. (2007).

Auditory-visual processing represented in the human superior temporal gyrus. *Neuroscience*, **145** (1), 162–184.

Ritter, W., Simson, R., & Vaughn, H. (1988). Effects of the amount of stimulus information processed on negative event-related potentials. *Electroencephalography and Clinical Neuropsychiology*, **28**, 244–258.

Rosenblum, L.D., Schmuckler, M.A., & Johnson, J.A. (1997). The McGurk effect in infants. *Perception & Psychophysics*, **59** (3), 347–357.

Ross, L.A., Molholm, S., Blanco, D., Gomez-Ramirez, M., Saint-Amour, D., & Foxe, J. (2011). The development of multisensory speech perception continues into the late childhood years. *European Journal of Neuroscience*, **33** (12), 2329–2337.

Skipper, J.I., Nusbaum, H.C., & Small, S.L. (2005). Listening to talking faces: motor cortical activation during speech perception. *NeuroImage*, **25**, 76–89.

Spreng, M. (1980). Influence of impulsive and fluctuating noise upon physiological excitations and short-time readaptation. *Scandanavian Audiology*, **12** (Suppl.), 299–306.

Stekelenburg, J.J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audio-visual events. *Journal of Cognitive Neuroscience*, **19** (12), 1964–1973.

Sumby, W., & Pollack, I. (1954). *Visual contribution to speech intelligibility in noise, Journal of the Acoustical Society of America*, **26**, 212–215.

Tanabe, H.C., Honda, M., & Sadato, N. (2005). Functionally segregated neural substrates for arbitrary audio-visual paired-association learning. *Journal of Neuroscience*, **25** (27), 6409–6418.

Teder-Salejarvi, W.A., McDonald, J.J., DiRusso, F., & Hillyard, S.A. (2002). An analysis of audio-visual crossmodal integration by means of event-related potential (ERP) recordings. *Cognitive Brain Research*, **14**, 106–114.

Teinonen, T., Aslin, R., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition*, **108**, 850–855.

Thomas, M.S.C., Annaz, D., Ansari, D., Serif, G., Jarrold, C., & Karmiloff-Smith, A. (2009). Using developmental trajectories to understand developmental disorders. *Journal of Speech, Language, and Hearing Research*, **52**, 336–358.

Thornton, A.R.D. (2008). Evaluation of a technique to measure latency jitter in event-related potentials. *Journal of Neuroscience Methods*, **8** (1), 248–255.

Tremblay, C., Champoux, F., Voss, P., Bacon, B.A., Lapore, F., & Theoret, H. (2007). Speech and non-speech audio-visual illusions: a developmental study. *PLoS*, **8**, 742.

Van Wassenhove, V., Grant, K.W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences, USA*, **102** (4), 1181–1186.

Viswanathan, D., & Jansen, B.H. (2010). The effect of stimulus expectancy on dishabituation of auditory evoked potentials. *International Journal of Psychophysiology*, **78**, 251–256.

Wada, Y., Shirai, N., Midorikawa, A., Kanazawa, S., Dan, I., & Yamaguchi, M.K. (2009). Sound enhances detection of visual target during infancy: a study using illusory contours. *Journal of Experimental Child Psychology*, **102** (3), 315–322.

Wightman, F., Kistler, D., & Brungart, D. (2006). Informational masking of speech in children: auditory–visual integration. *Journal of the Acoustical Society of America*, **119** (6), 3940–3949.

Wright, T.M., Pelphrey, K.A., Allison, T., McKeown, M.J., & McCarthy, G. (2003). Polysensory interactions along lateral temporal regions evoked by audio-visual speech. *Cerebral Cortex*, **13**, 1034–1043.