J. Child Lang., Page 1 of 31. © Cambridge University Press 2016
doi:10.1017/S030500091500077X

Children perceive speech onsets by ear and eye*

SUSAN JERGER

School of Behavioral and Brain Sciences, GR4·1, University of Texas at Dallas, and Callier Center for Communication Disorders, Richardson, Texas

MARKUS F. DAMIAN

School of Experimental Psychology, University of Bristol

NANCY TYE-MURRAY

Department of Otolaryngology–Head and Neck Surgery, Washington University School of Medicine

AND

HERVÉ ABDI

School of Behavioral and Brain Sciences, GR4·1, University of Texas at Dallas

(*Received* 1 *April* 2015 – *Revised* 11 *September* 2015 – *Accepted* 27 *November* 2015)

ABSTRACT

Adults use vision to perceive low-fidelity speech; yet how children acquire this ability is not well understood. The literature indicates that

[*] This research was supported by the National Institute on Deafness and Other Communication Disorders, grant DC-00421. Dr Abdi would like to acknowledge the support of an EURIAS fellowship at the Paris Institute for Advanced Studies (France), with the support of the European Union's 7th Framework Program for research, and funding from the French state managed by the *Agence Nationale de la Recherche* (program: *Investissements d'avenir, ANR-11-LABX-0027-01 Labex RFIEA+*). Sincere appreciation to (i) speech science colleagues for their guidance and advice to adopt a perceptual criterion for editing the non-intact stimuli and (ii) Dr Peter Assmann for generously giving of his time, talents, and software to prepare Figure 1. We thank Dr Brent Spehar for recording the audiovisual stimuli. We thank the children and parents who participated and the research staff who assisted, namely Aisha Aguilera, Carissa Dees, Nina Dinh, Nadia Dunkerton, Alycia Elkins, Brittany Hernandez, Cassandra Karl, Demi Krieger, Michelle McNeal, Jeffrey Okonye, Rachel Parra, and Kimberly Periman of UT-Dallas (data collection, analysis, presentation), and Derek Hammons and Scott Hawkins of UT-Dallas and Brent Spehar of Washington University School of Medicine (computer programming). Address for correspondence: Susan Jerger, School of Behavioral and Brain Sciences, GR4·1, University of Texas at Dallas, 800 W. Campbell Rd, Richardson, TX 75080. tel: 512-216-2961; e-mail: sjerger@utdallas.edu

children show reduced sensitivity to visual speech from kindergarten to adolescence. We hypothesized that this pattern reflects the effects of complex tasks and a growth period with harder-to-utilize cognitive resources, not lack of sensitivity. We investigated sensitivity to visual speech in children via the phonological priming produced by low-fidelity (non-intact onset) auditory speech presented audiovisually (see dynamic face articulate consonant/rhyme b/ag; hear non-intact onset/rhyme: –b/ag) vs. auditorily (see still face; hear exactly same auditory input). Audiovisual speech produced greater priming from four to fourteen years, indicating that visual speech filled in the non-intact auditory onsets. The influence of visual speech depended uniquely on phonology and speechreading. Children – like adults – perceive speech onsets multimodally. Findings are critical for incorporating visual speech into developmental theories of speech perception.

INTRODUCTION

In everyday conversations adults perceive speech by ear and eye, yet the development of this critical audiovisual property of speech perception is still not well understood. In fact, the extant child research reveals that – compared to adults – children exhibit reduced sensitivity to the articulatory gestures of talkers (i.e. visual speech). The McGurk task (McGurk & MacDonald, 1976) well illustrates this maturational difference in sensitivity to visual speech. In this task, individuals are presented with audiovisual stimuli with conflicting auditory and visual onsets (e.g. hear /ba/ and see /ga/). Whereas adults typically perceive a blend of the auditory and visual inputs (e.g. /da/ or /ða/) and rarely report perceiving the auditory /ba/, children, by contrast, report perceiving the /ba/ (auditory capture) 40% to 60% of the time (McGurk & MacDonald, 1976). Because visual speech plays a role in learning the phonological structure of spoken language (e.g. Locke, 1993; Mills, 1987), it is critical to understand how children utilize visual speech cues.

The influence of visual speech on children's audiovisual speech perception clearly increases with age, but the precise timecourse for achieving adultlike benefit from visual speech remains unclear. Numerous studies report that (i) children from roughly five through eleven years of age benefit less than adults from visual speech whereas (ii) adolescents (preteens–teenagers) show an adultlike visual speech advantage (e.g. Desjardins, Rogers & Werker, 1997; Dodd, 1977; Erdener & Burnham, 2013; Jerger, Damian, Spence, Tye-Murray & Abdi, 2009; McGurk & MacDonald, 1976; Ross, Molholm, Blanco, Gomez-Ramirez, Saint-Amour & Foxe, 2011; Tremblay, Champoux, Voss, Bacon, Lepore & Theoret, 2007; Wightman, Kistler & Brungart, 2006). Developmental improvements in sensitivity to visual speech have been attributed to changes in (i) the perceptual weights given to visual speech

(Green, 1998), (ii) articulatory proficiency and/or speechreading skills (e.g. Desjardins *et al*., 1997; Erdener & Burnham, 2013), and (iii) linguistic skills and language-specific tuning (Erdener & Burnham, 2013; Sekiyama & Burnham, 2004). Notable complications to this story are suggested, however, by several studies reporting significant sensitivity to visual speech in three- to five-year-olds (Holt, Kirk & Hay-McCutcheon, 2011; Lalonde & Holt, 2015), six- to seven-year-olds (Fort, Spinelli, Savariaux & Kandel, 2012), and eight-year-olds (Sekiyama & Burnham, 2004, 2008). Some of these studies stressed that performance in young children can be influenced by visual speech when the children are tested with developmentally appropriate measures and task demands. This viewpoint encourages us to consider the possible bases underlying children's developmental insensitivity to visual speech. Toward this end, Jerger *et al*. (2009) adopted a dynamic systems theoretical viewpoint (Smith & Thelen, 2003).

Dynamic systems theory

Dynamic systems theory proposes two relevant points for understanding the influence of visual speech in children: (i) multiple interactive factors form the basis of developmental change, and (ii) children's early skills are 'softly assembled' systems that reorganize into more mature, stable forms in response to environmental and internal forces (Smith & Thelen, 2003). Evoked potential studies support such a developmental reorganization and restructuring of the phonological system (Bonte & Blomert, 2004). During these developmental transitions, processing systems are less robust and children cannot easily use their cognitive resources; thus performance is less stable and more affected by methodological approaches and task demands (Evans, 2002). From this perspective, children's reduced sensitivity to visual speech may be incidental to developmental transformations, their processing by-products, and experimental contexts. Clearly, previous research has shown a greater influence of visual speech on children's performance when task demands were modified to be more child-appropriate (Desjardins *et al*., 1997; Lalonde & Holt, 2015). Further, sensitivity to visual speech has been shown to vary in the same children as a function of stimulus/task demands (Jerger, Damian, Tye-Murray & Abdi, 2014).

We propose that some experimental variables that might have contributed to children's reduced sensitivity to visual speech are the use of (i) complex tasks/audiovisual stimuli (e.g. targets embedded in noise or competing speech; McGurk stimuli with conflicting auditory and visual onsets) – because they make listening more challenging or less natural and familiar – and (ii) high-fidelity auditory speech – because it makes visual speech less relevant. The purpose of the present research was to evaluate whether sensitivity to visual speech in children might be increased by the use of stimuli with (i) congruent onsets that invoke more prototypical and representative audiovisual speech processes, and (ii) non-intact auditory onsets that increase the need for visual speech without involving noise. Below we briefly introduce our new stimuli and discuss the current task and its possible benefits for studying the influence of visual speech on performance by children.

Stimuli for the New Visual Speech Fill-In Effect

The new stimuli are words and nonwords with an intact consonant/rhyme in the visual track coupled to a non-intact onset/rhyme in the auditory track (our methodological criterion excised about 50 ms for words and 65 ms for nonwords; see 'Method'). Stimuli are presented in audiovisual vs. auditory modes. Example stimuli for the word *bag* are: (i) audiovisual: intact visual (b/ag) coupled to non-intact auditory (–*b*/ag) and (ii) auditory: static face coupled to the same non-intact auditory (–*b*/ag). Our idea was to insert visual speech into the 'nothingness' created by the excised auditory onset to study the possibility of a Visual Speech Fill-In Effect (Jerger *et al.*, 2014), which occurs when performance for the SAME auditory stimulus DIFFERS depending upon the presence/absence of visual speech. Responses illustrating a Visual Speech Fill-In Effect for a repetition task (Jerger *et al.*, 2014) are perceiving /bag/ in the audiovisual mode but /ag/ in the auditory mode. Below we overview our new approach – the multimodal picture–word task with low-fidelity speech (non-intact auditory onsets).

Multimodal picture–word task

In the widely used picture word interference task (Schriefers, Meyer & Levelt, 1990), participants name pictures while attempting to ignore nominally irrelevant speech distractors. Previous research (e.g. Jerger, Martin & Damian, 2002; Jerger *et al.*, 2009) has established that congruent onsets, such as [picture]–[distractor] pairs of [bug]–[bus], speed up picture naming times relative to neutral (or baseline) vowel onsets, such as [bug]–[onion]. A congruent onset is thought to prime picture naming because it creates crosstalk between the phonological representations that support speech production and perception (Levelt, Schriefers, Vorberg, Meyer, Pechmann & Havinga, 1991). Congruent distractors are assumed to spread activation from input to output phonological representations, a process fostering faster selection of speech segments during naming (Roelofs, 1997). Our 'multimodal' version of this task (Jerger *et al.*, 2009) administers audiovisual stimuli (Quicktime movie files). The to-be-named pictures appear on the T-shirt of a talker whose face moves (audiovisual speech utterance) or stays artificially still (auditory speech utterance coupled with still video). Hence, the speech distractors are presented audiovisually or auditorily only, a manipulation that enables us to study the influence of visual speech on phonological priming.

In a previous study with the multimodal picture–word task and high fidelity distractors (Jerger *et al.*, 2009), we observed a U-shaped developmental function with a significant influence of visual speech on phonological priming in four-year-olds and twelve-year-olds, but not in five- to nine-year-olds. Consistent with our dynamic systems theoretical viewpoint (Smith & Thelen, 2003), we proposed that phonological knowledge was reorganizing – particularly from five to nine years – into a more elaborated, systematized, and robust resource for supporting a wider range of activities, such as reading. The phonological knowledge supporting visual speech processing was not as readily accessed and/or retrieved during this pronounced period of restructuring for the reasons elaborated above (see also Jerger *et al.*, 2009). As noted above, our current research attempts to moderate these possible internal/external influences by using congruent audiovisual stimuli with non-intact auditory onsets. Our focus on speech onsets may be key because – relative to the other parts of an utterance – onsets are easier to speechread, more reliable with less articulatory variability, and more stressed (Gow, Melvold & Manuel, 1996). In two studies, we addressed research questions about the relation between phonological priming in the auditory vs. audiovisual modes as a function of the characteristics of the stimuli (Analysis 1) and the children's ages and verbal abilities (Analysis 2).

ANALYSIS 1: STIMULUS CHARACTERISTICS

The general aim of this analysis was to assess the influence of visual speech on phonological priming by high- vs. low-fidelity auditory speech in children from four to fourteen years. Whereas the auditory fidelity was manipulated from high to low (intact vs. non-intact onsets), the visual fidelity always remained high (intact). Primary research questions were whether – in all age groups – (i) the presence of visual speech would fill in the non-intact auditory onsets and prime picture naming more effectively than auditory speech alone and (ii) phonological priming would display a greater influence of visual speech for non-intact than intact auditory onsets. Finally, a secondary research question concerned LEXICAL STATUS, namely whether phonological priming in all age groups would display a greater influence of visual speech for nonwords than words (e.g. *baz* vs. *bag*). Some important qualities that may influence the effects of visual speech are: (i) CONGRUENT DIMENSIONS, (ii) INTEGRAL PROCESSING OF SPEECH CUES, and (iii) LOW-FIDELITY AUDITORY SPEECH.

STIMULUS CHARACTERISTICS AND PREDICTIONS

Congruent dimensions

Evidence suggests that audiovisual utterances with congruent rather than conflicting McGurk-like dimensions produce different perceptual experiences.

For example, Vatakis and Spence (2007) manipulated the temporal onsets of congruent vs. conflicting auditory and visual inputs and found that listeners were significantly less sensitive to temporal differences when onsets were congruent. Brain activation patterns also differ for congruent vs. conflicting audiovisual speech, with supra-additivity (greater than the sum of unimodal inputs) for the former but sub-additivity for the latter (Calvert, Campbell & Brammer, 2000). Congruent dimensions also possess lawful relatedness that produces strong cues that the auditory and visual inputs originated from the same speaker and should be integrated (Stevenson, Wallace & Altieri, 2014). Thus, in terms of a multisensory perceptual experience, congruent onsets offer some advantages compared to conflicting onsets. The data below also clearly indicate that the speech cues of consonant–vowel stimuli are processed integrally.

Integrality of speech cues

To study the integrality of speech cues, the Garner task (1974) requires participants to (i) attend selectively to a target cue such as a consonant (e.g. /b/ vs. /g/) and (ii) try to ignore a non-target cue such as a vowel that is held constant (/ba/ vs. /ga/) or varies irrelevantly (/ba/, /bi/ vs. /ga/, /gi/). Results have shown that irrelevant variation in the vowels interferes with classifying the consonants and vice versa (e.g. Tomiak, Mullennix & Sawusch, 1987). Green and Kuhl (1989) established that this tight coupling between auditory speech cues extends to audiovisual speech cues. All these results indicate that listeners cannot ignore one speech cue and selectively attend to another. Instead, listeners perceive the cues integrally. Results on the Garner task imply that our auditory and visual speech onsets should be processed integrally.

Low-fidelity (non-Intact) auditory speech

The literature shows a shift in the relative weights of the auditory and visual modes as the quality of the inputs shifts. To illustrate: when listening to McGurk stimuli with degraded auditory speech, children with normal hearing respond more on the basis of the intact visual input (Huyse, Berthommier & Leybaert, 2013). When the visual input is also degraded, however, the children respond more on the basis of the degraded auditory input. Children with normal hearing or mild–moderate hearing loss and good auditory word recognition – when listening to conflicting inputs such as auditory /meat/ coupled with visual /street/ – respond on the basis of the auditory input (Seewald, Ross, Giolas & Yonovitz, 1985). In contrast, children with more severe hearing loss – and more degraded perception of auditory input – respond more on the basis of the visual input. Finally, when Japanese individuals listen to high-fidelity auditory input, they do not show a McGurk effect; but when they listen to degraded auditory input, they do show the effect (Sekiyama & Burnham, 2008; Sekiyama & Tohkura, 1991). These results indicate that the relative weighting of auditory and visual speech is modulated by the relative quality of each input. Recent neuroscience studies also support this differential weighting, as they reveal that the functional connectivity between the auditory and visual cortices and the superior temporal sulcus (STS, an area of audiovisual integration) changes with input fidelity, with increased connectivity between the STS and the sensory cortex with the higher-fidelity input (Nath & Beauchamp, 2011).

In short, our auditory and visual speech cues are congruent and should be processed in an integral manner. The auditory and visual speech inputs should be weighed differentially depending on the quality of the auditory input. Thus we predict that (i) visual speech will fill in the non-intact auditory onsets and prime picture naming more effectively than auditory speech alone, and (ii) children will be more sensitive to visual speech for non-intact than intact auditory input. In addition to our primary research questions, a secondary research question evaluated whether lexical status affects children's sensitivity to visual speech.

LEXICAL STATUS AND PREDICTIONS

The literature contrasting the McGurk effect for words vs. nonwords indicates that the McGurk effect occurs for both types of stimuli. Within this evidence, some results have revealed that lexical status impacts the McGurk effect. For example, visual speech influences listeners more often when (i) stimuli are words rather than nonwords (Barutchu, Crewther, Kiely, Murphy & Crewther, 2008) or (ii) the visual input forms a word and the auditory input forms a nonword (Brancazio, 2004). By contrast, however, other results have shown a strong McGurk effect for both nonwords and words, with performance not appearing to be influenced by meaningfulness (Sams, Manninen, Surakka, Helin & Katto, 1998). With regard to studies assessing the McGurk effect with only word stimuli in isolation, one study (Dekle, Fowler & Funnell, 1992) observed a strong McGurk effect whereas the other study (Easton & Basala, 1982) reported no visual influence on performance. In short, these studies do not provide consistent results or predictions

In contrast to the mixed results summarized above, the hierarchical model of speech segmentation (Mattys, White & Melhorn, 2005) provides unambiguous predictions for words vs. nonwords. The model proposes that listeners assign the greatest weight to lexical–semantic content when listening to words. If the lexical–semantic content is compromised, however, listeners assign the greatest weight to phonetic–phonological content. If both the lexical–semantic and phonetic–phonological content are compromised, listeners assign the greatest weight to acoustic–temporal content. It is also assumed that monosyllabic words such as our stimuli (*bag*) may activate their lexical representations without requiring phonological decomposition whereas nonwords (*baz*) require phonological decomposition (Mattys, 2014).

If these ideas generalize to our task, word stimuli should be heavily weighted in terms of lexical–semantic content but nonword stimuli should be heavily weighted in terms of phonetic–phonological content for both the audiovisual and auditory modes. We predict that children's sensitivity to visual speech will vary depending on the relative weighting and decomposition of the phonetic–phonological content. To the extent that a greater weight on phonetics–phonology increases children's awareness of the phonetic–phonological content and visual speech phonetic cues, we predict that children will show a significantly greater influence of visual speech relative to auditory speech for nonwords than for words. In agreement with Campbell (1988), we view visual speech as an extra phonetic resource that adds another type of phonetic feature.

Although we critically evaluate the influence of child factors in Analysis 2, we plot results as a function of age in Analysis 1. To briefly address age, the literature reviewed above predicts that – although benefit from visual speech improves with age – children relative to adults show significantly reduced benefit up to the adolescent years. We have argued above, however, that performance for our non-intact stimuli will reveal MORE sensitivity to visual speech. We thus predict that phonological priming effects will show influences of visual speech from four to fourteen years.

METHOD

Participants

Participants were 132 native English-speaking children ranging in age from 4;2 to 14;5 (55% boys). The racial distribution was 70% White, 13% Asian, 11% Black, and 6% Multiracial, with 9% reporting Hispanic ethnicity. Participants had normal (age-based when appropriate) hearing sensitivity, visual acuity (including corrected to normal), auditory word recognition (Ross & Lerman, 1971), articulatory proficiency (Goldman & Fristoe, 2000), and visual perception (Beery & Beery, 2004). Participants were divided into four age groups (30 to 38 children each) based on chronological age (four- to five-year-olds: $M = 4;11$, $SD = 0{\cdot}53$; six- to seven-year-olds: $M = 7;00$, $SD = 0{\cdot}59$; eight- to ten-year-olds: $M = 9;02$, $SD = 0{\cdot}87$; and eleven- to fourteen-year-olds: $M = 12;04$, $SD = 1{\cdot}24$). These groups will be referred to as five-year-olds, seven-year-olds, nine-year-olds, and twelve-year-olds. Details for the groups are presented in Analysis 2. Participants accurately pronounced the onsets of the pictures' names; the offsets were also accurately pronounced except for three five-year-olds (who substituted /θ/ for /s/ in *gas* and *geese* or omitted /t/ in *ghost*). Two five-year-olds had to be taught the names of some pictures (*geese, beads*, and/or *gun*). To ensure that the experimental results were reflecting performance for words vs. nonwords, participants' knowledge of the word distractors was tested by parental report and a picture-pointing task. Thirty-one children had to be taught the meaning of a distractor; the mean number of unknown distractors averaged 0·917 in the five-year-olds, 0·414 in the seven-year-olds, and 0·016 in the nine- to twelve-year-olds. Mean naming times for the taught vs. previously known words did not differ; no trials were eliminated.

Materials and instrumentation: picture–word task

Pictures and distractors. The entire set of materials consisted of experimental items (8 pictures and 12 distractors) and filler items (16 pictures and 16 distractors). The experimental pictures and phonologically related distractors were words/nonwords beginning with the consonants /b/ or /g/ coupled with the vowels /i/, /æ/, /ʌ/, or /o/. The baseline distractors were words/nonwords beginning with the vowels /i/, /æ/, /ʌ/, or /o/. Illustrative items for the picture [bug] are [picture]–[word/nonword] pairs of [bug]–[bus/buv] for the phonologically related condition and [bug]–[onion/onyit] for the baseline condition (see 'Appendix A' for items, available at http://dx.doi.org/10.1017/S030500091500077X). The word and nonword distractors were constructed to have as comparable phonotactic probabilities as possible. In brief, the positional segment frequencies for the words vs. nonwords averaged respectively ·1593 vs. ·1570 (adult values) and ·1911 vs. ·1805 (child values); the biphone frequencies averaged ·0050 vs. ·0056 (adult values) and ·0071 vs. ·0074 (child values) (Storkel & Hoover, 2010; Vitevitch & Luce, 2004; see Jerger *et al*., 2014, for details). The filler items were pictures and word/nonword distractors NOT beginning with /b/ or /g/. Illustrative filler items are the [picture]–[word/nonword] pairs of [dog]–[cheese/cheeg], [shirt]–[pickle/pimmel], and [cookies]–[horse/hork]. To emphasize the distinctiveness between the words and nonwords, if a filler item (e.g. [dog]–[cheese]) was used for the words, its counterpart (e.g. [dog]–[cheeg]) was not used for the nonwords and vice versa. This strategy yielded 8 different picture–distractor filler items each for the words and the nonwords.

Stimulus preparation. The distractors were recorded at the Audiovisual Recording Lab, Washington University School of Medicine. The talker was an eleven-year-old boy actor with clearly intelligible speech. His full facial image and upper chest were recorded. He started and ended each utterance with a neutral face / closed mouth. The color video signal was digitized at 30 frames/s with 24-bit resolution at a 720 × 480 pixel size. The auditory signal was digitized at 48 kHz sampling rate with 16-bit amplitude resolution. The

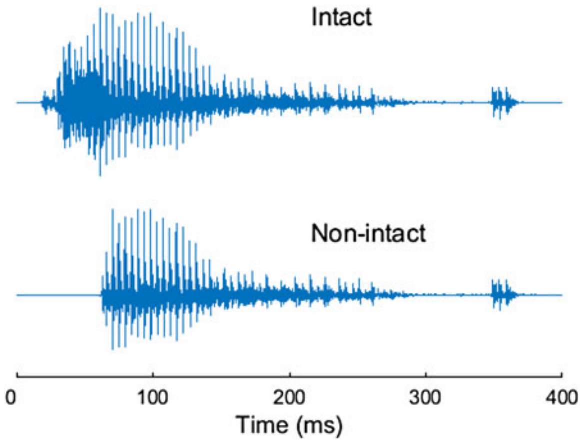

Fig. 1. Display of the intact vs. non-intact auditory waveforms for the word *bag*.

utterances were adjusted to equivalent A-weighted root mean square sound levels. The video track was routed to a high-resolution monitor, and the auditory track was routed through a speech audiometer to a loudspeaker. The intensity level of the distractors was approximately 70 dB SPL. The to-be-named colored pictures were scanned into a computer as 8-bit PICT files and edited to achieve objects of a similar size on a white background.

Editing the auditory onsets. We edited the auditory track of the phonologically related distracters by locating the /b/ or /g/ onsets visually and auditorily with Adobe Premiere Pro and Soundbooth (Adobe Systems Inc., San Jose, CA) and loudspeakers. We applied a perceptual criterion to operationally define a non-intact onset. We excised the waveform in 1 ms steps from the identified auditory onset (first deviations from baseline) to the point in the later waveform for which at least four of five trained listeners heard the vowel as the onset (auditory mode). This process removed the excised portion of the acoustic signal and left the alignment between the auditory and visual tracks as originally produced by the speaker. Splice points were always at zero axis crossings. Using our perceptual criterion, we excised on average 52 ms (/b/) and 50 ms (/g/) from the word onsets and 63 ms (/b/) and 72 ms (/g/) from the nonword onsets. Figure 1 displays the intact vs. non-intact waveforms for the word *bag*.

We next formed audiovisual (dynamic face) and auditory (static face) modes of presentation for the stimuli. In our experimental design, the auditory mode controls for the influence on performance of any remaining coarticulatory cues in the input. More specifically, we compare results for the non-intact stimuli in the auditory vs. audiovisual modes. Any coarticulatory cues in the auditory input are held constant in the two modes. Thus any influence on picture naming due to articulatory cues should be controlled, and this should allow us to evaluate whether the addition of visual speech influences performance.

Audiovisual and auditory modes. Stimuli were Quicktime movie files. For the audiovisual mode, the children saw (i) 924 ms (experimental trials) or 627 or 1,221 ms (filler-item trials) of the talker's still face and upper chest, followed by (ii) an audiovisual utterance of one distractor and the presentation of one picture on the talker's T-shirt five frames before the auditory onset of the utterance (auditory distractor lags picture), followed by (iii) 924 ms of still face and picture. For the auditory mode, the child heard the same event but the video track was edited to contain only the talker's still face. The onset of the picture occurred in the same frame for the intact and non-intact distracters. The relationship between the onsets of the picture and the distractor, termed stimulus onset asynchrony (SOA), must also be considered for the picture–word task.

SOA. Phonologically related distracters typically produce a maximal effect on naming when the onset of the auditory distractor lags the onset of the picture with a SOA of about 150 ms (Damian & Martin, 1999; Schriefers *et al*., 1990). Our SOA was five frames or about 165 ms (frame size of 33 ms) as used previously (Jerger *et al*., 2009). Because the picture remained in the same frame for the intact and non-intact stimuli, however, the auditory non-intact onset altered the target SOA of 165 ms and the natural temporal synchrony between the visual and auditory speech onsets. Below we consider these issues.

With regard to altering the SOA, the child literature does not provide evidence about whether the slight temporal shift in the SOA produced by the non-intact onset affects picture naming results. Our experimental design, however, should provide data that can control for this issue. To do so, we will compare results for the non-intact stimuli in the auditory vs. audiovisual modes. The shift in the auditory onset is held constant in the two modes; thus any influence on picture naming due to the shift in the auditory onset should be controlled. This should allow us to evaluate whether the addition of visual speech influences performance.

With regard to altering the temporal synchrony between modes, visual speech normally leads auditory speech (Bell-Berti & Harris, 1981), but the degree to which visual speech leads varies appreciably (ten Oever, Sack, Wheat, Bien & van Atteveldt, 2013). Thus listeners are accustomed to natural variability in this asynchrony. Adults synthesize visual and auditory speech into a single multisensory event – without any detection of the asynchrony or any effect on intelligibility – when the visual speech leads the auditory speech by as much as 200 ms (Grant, van Wassenhove & Poeppel, 2004). Detecting asynchrony between audiovisual speech inputs (simultaneity judgments) is similar in adults and ten- to eleven-year-olds when visual speech leads (Hillock, Powers & Wallace, 2011). This evidence suggests that the alternation in the SOA produced by the non-intact onsets will not affect the children's assimilation of an audiovisual distractor into a single multisensory event. Below we summarize our final set of materials.

Final set of items. We administered two presentations of each experimental item (i.e. baseline, intact, and non-intact distractors) in the audiovisual and auditory modes. The items were randomly intermixed with the filler items in each mode and formed into four lists (which were presented forward or backward for eight variations). Each list contained 24 experimental (57%) and 18 filler-item (43%) trials. The items comprising a list varied randomly under the constraints that (i) no onset could repeat, (ii) the intact and non-intact pairs (e.g. bag and /–*b*/ag) could not occur without at least two intervening items, (iii) a non-intact onset must be followed by an intact onset, (iv) the mode must alternate after three repetitions, and (v) all types of onsets (vowel, intact /b/ and /g/, non-intact /b/ and /g/, and not /b/ or /g/) must be dispersed uniformly throughout the lists. The presentation of items was counterbalanced so that 50% of items occurred first in the auditory mode and 50% occurred first in the audiovisual mode. The number of intervening items between the intact vs. non-intact pairs (and vice versa) averaged ten items.

Naming responses. Participants named pictures by speaking into a unidirectional microphone mounted on an adjustable stand. The utterances were digitally recorded. To quantify naming times, the computer triggered a counter/timer (resolution less than one ms) at the initiation of a movie file. The timer was stopped by the onset of the participant's vocal response into the microphone, which was fed through a stereo mixing console amplifier and 1 dB step attenuator to a voice-operated relay (VOR). A pulse from the VOR stopped the timing board via a data module board. If necessary, the participant's speaking level, the position of the microphone or child, and/or the setting on the 1 dB step attenuator were adjusted to ensure that the VOR triggered reliably. The counter timer values were corrected for the amount of silence in each movie file before the onset of the picture.

Procedure

The children completed the multimodal picture–word task along with other procedures in three sessions, scheduled approximately ten days apart. The order of presentation of the word vs. nonword conditions was counterbalanced across participants in each age group. Results were collapsed across the counterbalancing conditions. In the first session, the children completed three of the word (or nonword) lists; in the second session, the children completed the fourth word (or nonword) list and the first nonword (or word) list; and in the third session, the children completed the remaining three nonword (or word) lists. Individual lists were administered in separated listening conditions. A variable number of practice trials preceded the presentation of each list.

At the start of the first session, a tester showed each picture on a 5" *x* 5" card and asked the participant to name the picture; the tester taught the target names of any pictures named incorrectly. Next the tester flashed some picture cards quickly and modeled speeded naming. The child copied the tester. Speeded naming practice trials went back and forth between tester and child until the child was naming the pictures fluently. Mini-practice trials started each of the other sessions.

For formal testing, a tester sat at a computer workstation and initiated each trial by pressing a touch pad (out of child's sight). The children, with a co-tester alongside, sat at a distance of 71 cm directly in front of an adjustable height table containing the computer monitor and loudspeaker. Trials that the co-tester judged flawed (e.g. child squirmed out of position, child triggered microphone with non-speech) were deleted online and re-administered after intervening items. The children were told they would see and hear a boy whose mouth would sometimes be moving and sometimes not. For the words, participants were told that they might hear words or nonwords; for the nonwords, participants were told that they would always hear nonwords. We emphasized that the talking was not important. Participants were told to focus only on (i) watching for a picture that would pop up on the boy's T-shirt and (ii) naming it as quickly and as accurately as possible. The participant's view of the picture subtended a visual angle of 5·65° vertically and 10·25° horizontally; the view of the talker's face subtended a visual angle of 7·17° vertically (eyebrow – chin) and 10·71° horizontally (eye level). Finally, participants also completed an explicit repetition task (always presented after the completion of the picture–word task) to assess the perception of the distractor onsets.

RESULTS

Preliminary analyses

'Appendix B' (available at http://dx.doi.org/10.1017/S030500091500077X) details (i) the accuracy of perceiving the onsets and (ii) the quality of the picture–word data (e.g. number of missing trials). In addition to these results, we analyzed the picture–word data preliminarily to determine whether results could be collapsed across the different distractor onsets (/b/ vs. /g/). Appendix C (available at http://dx.doi.org/10.1017/S030500091500077X) details these results. Briefly, separate factorial mixed-design analyses of variance were performed for the baseline and phonologically related distractors. Findings

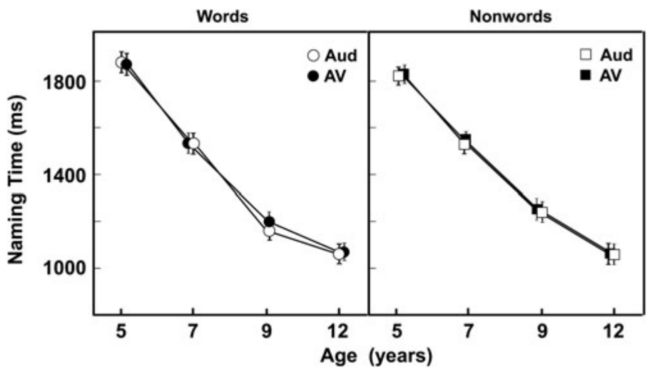

Fig. 2. Average picture naming times for the age groups in the presence of the vowel-onset baseline distractors presented in the auditory (Aud) or audiovisual (AV) modes for the words (left) and nonwords (right). Error bars are standard errors of the mean. Each age group represents a range of chronological ages (see text).

indicated that the different onsets influenced results for the phonologically related distractors but not for the baseline distractors. Specifically, overall picture naming speed was facilitated slightly more for the /b/ than /g/ onset (−147 vs. −117 ms). The effect of the onsets was also slightly more pronounced for the audiovisual than auditory mode (38 vs. 20 ms).

Despite these statistically significant outcomes, the differences in performance due to onset were small and did not interact with lexical status (words vs. nonwords) or fidelity (intact vs. non-intact). Thus, we developed a dual-pronged approach. For the primary analyses below, naming times were collapsed across the onsets to make the principal story clearer. For one key analysis with the collapsed onsets, however (determining whether/how visual speech influenced performance by assessing the difference between each pair of audiovisual–auditory naming times), the analysis was repeated separately for the individual /b/ and /g/ onsets. This analysis provides strong evidence for readers interested in whether/how the speechreadability of the onsets influenced phonological priming (e.g. the bilabial /b/ is easier to speechread than the velar /g/; Tye-Murray, 2014).

Baseline picture–word naming times

Figure 2 shows average picture naming times for the age groups in the presence of the vowel-onset baseline distractors presented in the auditory or audiovisual modes for the words (left) and nonwords (right). Results were analyzed with a factorial mixed-design analysis of variance with one between-participants factor (four age groups) and two within-participant factors (lexical status [words vs. nonwords] and mode [auditory vs. audiovisual]). Results indicated that picture naming times decreased significantly as age increased ($F(3,128) =$

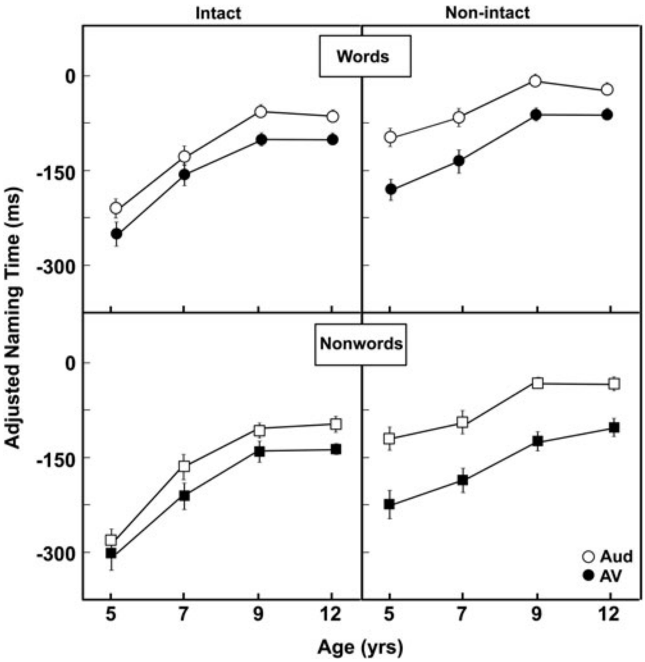

Fig. 3. Adjusted naming times in the age groups for the words and nonwords (lexical status: top vs. bottom panels) with intact and non-intact onsets (high vs. low fidelity: left vs. right panels) presented in the auditory (Aud) and audiovisual (AV) modes. The zero of the ordinate represents naming times for the baseline distractors (Fig. 1). Error bars are standard errors of the mean. Each age group represents a range of chronological ages (see text).

86·33, $MSE = 197462{\cdot}74$, $p < {\cdot}001$, partial $\eta^2 = {\cdot}669$). No other significant effect was observed. Picture naming times declined from about 1855 ms in the five-year-olds to 1065 ms in the twelve-year-olds for both words and nonwords in both modes. This finding agrees previous findings (e.g. Brooks & MacWhinney, 2000; Jerger *et al.*, 2002).

Phonologically related picture–word naming times

We quantified the priming produced by the phonologically related distractors on picture naming with adjusted naming times, derived by subtracting each participant's baseline naming times from his or her phonologically related naming times as in previous studies (e.g. Jerger *et al.*, 2009). Figure 3 depicts the adjusted naming times in the age groups for words and nonwords (top vs. bottom panels) in the auditory and audiovisual modes. Performance is shown for both the intact and non-intact stimuli (left vs. right panels).

TABLE 1. *Summary of statistical results*

A. Untransformed data: the dependent variable is adjusted naming times (naming time for phonologically related distractors minus naming time for baseline distractors).

Factors	*Mean square error*	*F value*	*p value*	*partial* η^2
Age Group	**34368·94**	**30·40**	**< ·001**	**.416**
Lexical Status	**16696·22**	**25·97**	**< ·001**	**·169**
Fidelity	**4324·70**	**199·55**	**< ·001**	**·609**
Mode	**4186·19**	**189·54**	**< ·001**	**·597**
Age Group × Fidelity	**4324·70**	**16·41**	**< ·001**	**·278**
Mode × Fidelity	**1421·23**	**78·80**	**< ·001**	**·381**
Mode × Lexical Status	**4456·09**	**3·97**	**·048**	**·030**
Lexical Status × Fidelity	2201·67	3·80	ns	·028
Age Group × Lexical Status	16696·22	0·11	ns	·003
Age Group × Mode	4186·19	0·59	ns	·014
Mode × Fidelity × Lexical Status	**2299·80**	**4·491**	**·035**	**·034**
Age Group × Lexical Status × Fidelity	2201·67	1·192	ns	·027
Age Group × Mode × Lexical Status	4456·09	0·51	ns	·011
Age Group × Mode × Fidelity	1421·23	2·51	ns	·055
Age Group × Mode × Lexical Status × Fidelity	2299·80	0·65	ns	·015

Results were analyzed with a factorial mixed-design analysis of variance with one between-participants factor (four age groups) and three within-participant factors (lexical status [words vs. nonwords], fidelity [intact vs. non-intact], and mode [auditory vs. audiovisual]). Table 1A summarizes the results (significant results are bolded). All four main factors significantly influenced how the phonologically related distractors primed OVERALL picture naming times, with an effect of (i) AGE GROUP, showing greater priming in the younger than the older children [five-year-olds: –208 ms; seven-year-olds: –143 ms; nine-year-olds and twelve-year-olds: –80 ms], (ii) LEXICAL STATUS, showing greater priming from the nonword than the word distractors [respectively –153 ms vs. –112 ms], (iii) FIDELITY, showing greater priming from the intact than the non-intact distractors [respectively –162 ms vs. –102 ms], and (iv) MODE, showing greater priming from the audiovisual than the auditory distractors [respectively –160 ms vs. –104 ms]. The significantly greater priming for the audiovisual mode is particularly relevant because this pattern highlights a significant influence of visual speech on performance.

A few interactions were also significant, but only one involved age group, namely an AGE GROUP X FIDELITY interaction (see Table 1A). As shown in Figure 3 and noted above, the intact (high-fidelity) distractors primed OVERALL picture naming more effectively than the non-intact (low-fidelity)

B. Proportion transformed data: the dependent variable is proportion derived by dividing adjusted naming time by baseline naming time.

Factors	*Mean square error*	*F value*	*p value*	*partial η^2*
Age Group	**·012**	**8·93**	**< ·001**	**·173**
Lexical Status	**·005**	**41·33**	**< ·001**	**·244**
Fidelity	**·001**	**266·03**	**< ·001**	**·369**
Mode	**·002**	**264·88**	**< ·001**	**·675**
Age Group × Fidelity	**·001**	**6·09**	**< ·001**	**·124**
Mode × Fidelity	**·001**	**86·38**	**< ·001**	**·400**
Mode × Lexical Status	**·001**	**6·79**	**·010**	**·051**
Lexical Status × Fidelity	**·001**	**4·42**	**·038**	**·035**
Age Group × Lexical Status	005	0·12	ns	·003
Age Group × Mode	·002	1·34	ns	·030
Mode × Fidelity × Lexical Status	**·001**	**7·46**	**·007**	**·055**
Age Group × Lexical Status × Fidelity	·001	0·71	ns	·018
Age Group × Mode × Lexical Status	·001	1·05	ns	·026
Age Group × Mode × Fidelity	·001	0·27	ns	·011
Age Group × Mode × Lexical Status × Fidelity	·001	0·77	ns	·019

NOTE: ns = p > ·05. Results of a mixed-design analysis of variance with one between-participants factor (Four Age Groups) and three within-participants factors (Lexical Status: word vs. nonword; Fidelity: intact vs. non-intact; Mode: auditory vs. audiovisual). The degrees of freedom are 1,128 for all factors except those involving Age Group wherein the degrees of freedom are 3,128.

distractors (compare right vs. left panels collapsed across mode and lexical status). This interaction arose because the relative effectiveness of the intact vs. non-intact distractors differed more in the five-year-olds (−104 ms) than in the older groups (seven-year-olds: −44 ms; nine-year-olds: −43 ms; twelve-year-olds: −39 ms). The other significant interactions (two-way and three-way) shown in Table 1A involved mode. To clarify these interactions – and determine whether visual speech significantly influenced performance – we quantified the difference between each pair (audiovisual–auditory) of adjusted naming times. For the sake of simplicity, we labeled all of the difference scores, for both the intact (high-fidelity) and non-intact (low-fidelity) stimuli, a VISUAL SPEECH EFFECT (VSPE) for these analyses. We should emphasize, however, that this VSPE is reflecting an actual filling in of some missing auditory cues for non-intact speech and, by contrast, an augmenting of auditory cues for intact speech. The difference scores are plotted in Figure 4 and represent the difference between the lines in Figure 3. The error bars show the 95% confidence intervals for the difference scores. Note that the confidence intervals do not provide relevant information about the intact and non-intact

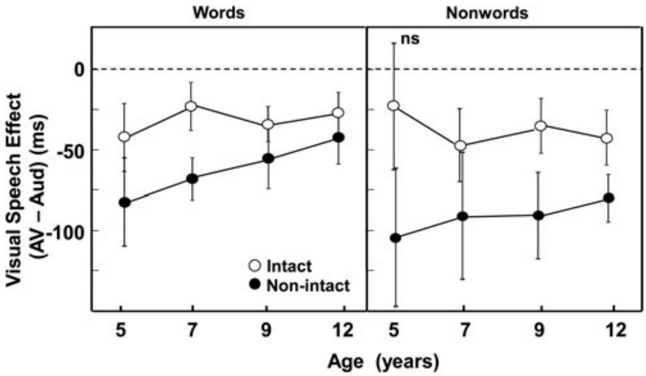

Fig. 4. Visual Speech Effect (VSPE; defined by the mean difference between audiovisual–auditory [AV–Aud] adjusted naming times) for the intact and non-intact onsets (high vs. low fidelity) of the words and nonwords (lexical status: left vs. right) in the age groups. The VSPE is reflecting an actual filling in of some missing auditory cues for non-intact speech and, by contrast, an augmenting of auditory cues for intact speech. Error bars show 95% confidence intervals. ALL datapoints showed significantly greater priming for the AV than Aud mode excepting one: nonwords intact, five-year-olds. Each age group represents a range of chronological ages (see text). ns = not significant.

conditions because only difference scores are interpretable for factors that are not independent.

The higher order (MODE X FIDELITY X LEXICAL STATUS) interaction occurred because the VSPE for the non-intact onsets (Figure 4 collapsed across age groups) was greater for the nonwords than the words (i.e. respectively 91 ms vs. 62 ms; left vs. right panels) whereas the VSPE for the intact onsets did not differ for the nonwords vs words (i.e. respectively 36 ms vs. 33 ms). Although this higher-order interaction may limit the interpretation of the lower-order interactions, we should nonetheless acknowledge the interactions between mode vs. fidelity and vs. lexical status. The MODE X FIDELITY interaction occurred because results showed a greater VSPE for the non-intact than intact onsets (respectively –77 ms vs. –34 ms; Figure 4 collapsed across age groups and lexical status). The MODE X LEXICAL STATUS interaction emerged because results showed a larger VSPE for nonwords than words (respectively –63 ms vs. –47 ms; Figure 4 collapsed across age groups and fidelity).

With regard to whether visual speech significantly influenced performance, the confidence intervals (Figure 4) address whether a given group showed a significant VSPE (i.e. did each result differ significantly from zero?). If the 95% confidence interval, or the range of plausible difference scores, does not contain zero, then the results are significant. The confidence intervals revealed a significant VSPE for all the non-intact

TABLE 2. *Confidence intervals (95%) for the adjusted naming times in Figure 2.*

	High fidelity (intact)		Low fidelity (non-intact)	
Mode		**Words**		
Age groups				
Auditory				
5	−240, −178	*	−129, −70	*
7	−165, −97	*	−97, −40	*
9	−77, −41	*	−29, +7	ns
12	−87, −51	*	−39, −9	*
Audiovisual				
5	−291, −211	*	−215, −146	*
7	−192, −116	*	−166, −105	*
9	−117, −76	*	−88, −45	*
12	−114, −77	*	−82, −49	*
		Nonwords		
Auditory				
5	−319, −240	*	−158, −83	*
7	−205, −122	*	−132, −57	*
9	−132, −82	*	−50, −16	*
12	−123, −69	*	−62, −14	*
Audiovisual				
5	−357, −246	*	−268, −178	*
7	−253, −167	*	−224, −145	*
9	−175, −107	*	−154, −91	*
12	−168, −108	*	−145, −88	*

NOTE: * = significant priming; ns = no priming; each age group represents a range of chronological ages (see text).

and intact onsets excepting one, namely intact nonwords in the five-year-olds.

Finally, confidence intervals for the results in Figure 3 are also of interest in terms of whether the phonologically related distractors significantly primed naming in each group. Our specific question was whether each adjusted naming time (difference score between phonologically related naming time and baseline naming time) in each group for each mode differed significantly from zero. Table 2 shows the 95% confidence intervals. Results indicated significant priming – the confidence interval did not contain zero – for all datapoints in Figure 3 excepting one; namely non-intact words, auditory mode in the nine-year-olds. Although values outside of 95% confidence intervals are relatively implausible, the lower limits neared zero for two significant results – non-intact nonwords, auditory mode in the nine-year-olds and twelve-year-olds – a pattern suggesting that we should have a lesser degree of confidence in the repeatability of these two outcomes.

With regard to the above effects of age, a complication is that the differences in the baseline naming times muddle an unequivocal interpretation of the results. In other words, the greater priming effects in the five-year-olds (Figure 3) could be a result of age or of these children's slower baseline naming times. A straightforward approach to controlling the baseline differences (see Damian & Dumay, 2007) is to develop priming proportions. Thus we divided each participant's adjusted naming times by her or his corresponding baseline naming times (i.e. [mean time in the phonologically related condition minus mean time in the baseline condition] divided by [mean time in the baseline condition]). A factorial mixed-design analysis of variance on these transformed data, with the same between- and within-participant factors, yielded the same pattern of results as above (see Table 1B). We continued to observe the significant effect of (i) AGE GROUP, showing greater priming in the younger than older children [five-year-olds: –0·110; seven-year-olds: –0·090; nine-year-olds and twelve-year-olds: –0·070], and the one age group interaction, AGE GROUP X FIDELITY, which was elaborated above.

With regard to the interactions that the VSPE clarified in Figure 4, the transformed data also continued to reveal the significant higher-order interaction (MODE X FIDELITY X LEXICAL STATUS) and the two lower-order interactions (MODE X FIDELITY and MODE X LEXICAL STATUS). A third lower-order interaction (LEXICAL STATUS X FIDELITY) also achieved significance ($p = {\cdot}038$). This interaction occurred because the difference between priming for the intact vs. non-intact stimuli was slightly greater for nonwords than words, with difference scores respectively of ·043 and ·036 for the proportion transformed data (and 66 vs. 53 ms for the untransformed data).

Finally, it is of interest to ask whether there was a complete or partial Visual Speech Fill-In Effect. The previous MODE X FIDELITY interaction indicates that phonological priming by the intact vs. non-intact distractors differed more for the auditory (−145 ms vs. −64 ms) than audiovisual (−179 ms vs. −141 ms) mode (see Figure 3). Clearly this interaction reflects a robust Visual Speech Fill-In Effect or, as indicated previously, a greater VSPE for the non-intact than intact onsets. However, the current question is whether the Visual Speech Fill-In Effect was complete or partial (in other words, were the non-intact audiovisual distractors as phonologically effective as their intact counterparts).

To evaluate whether phonological priming differed for the non-intact vs. intact audiovisual distractors, we carried out orthogonal contrasts (Abdi & Williams, 2010) on the mean audiovisual adjusted naming times collapsed across the words and nonwords. We found significantly greater priming from the intact than non-intact audiovisual distractors in all age groups: (*five-year-olds*, F_{contrast} (1,128) = 64·08, $MSE = 1421{\cdot}23$, $p < {\cdot}001$, partial $\eta^2 = {\cdot}334$; *seven-year-olds*, F_{contrast} (1,128) = 5·75, $MSE = 1421{\cdot}23$, $p = {\cdot}02$, partial $\eta^2 = {\cdot}043$; *nine-year-olds*, F_{contrast} (1,128) = 6·80, $MSE = 1421{\cdot}23$,

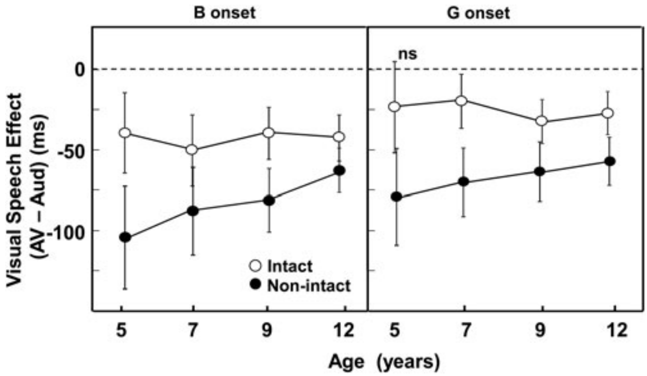

Fig. 5. Visual Speech Effect (VSPE; defined by the mean difference between audiovisual–auditory [AV–Aud] adjusted naming times) for the /b/ and /g/ onsets of the intact and non-intact inputs (high vs. low fidelity) in the age groups (results are collapsed across words and nonwords). The VSPE is reflecting an actual filling in of some missing auditory cues for non-intact speech and, by contrast, an augmenting of auditory cues for intact speech. Error bars show 95% confidence intervals. ALL datapoints showed significantly greater priming for the AV than Aud mode excepting one: /g/ onset, intact, five-year-olds. Each age group represents a range of chronological ages (see text). ns = not significant.

$p = \cdot 01$, partial $\eta^2 = \cdot 050$; *twelve-year-olds*, $F_{\text{contrast}}\ (1{,}128) = 7{\cdot}77$, $MSE = 1421{\cdot}23$, $p = \cdot 006$, partial $\eta^2 = \cdot 057$). Thus even though the Visual Speech Fill-In Effect was robustly effective, the non-intact audiovisual distractors were not as phonologically compelling as their intact counterparts.

VSPE for the individual /b/ and /g/ onsets. To probe the influence of visual speech as a function of the speechreadability of the onsets, we analyzed the VSPE scores – without collapsing across the onsets – with a factorial mixed-design analysis of variance with one between-participants factor (four age groups) and three within-participant factors (lexical status [words vs. nonwords], fidelity [intact vs. non-intact], and onset [b vs. g]). There was no significant effect of lexical status nor were there any interactions between lexical status and fidelity or onset; thus to graph the results, the VSPE for the onsets was collapsed across words and nonwords. Figure 5 portrays the collapsed VSPE for the /b/ and /g/ onsets in the high- (intact) and low- (non-intact) fidelity conditions in the age groups, along with the 95% confidence intervals.

The statistical analysis revealed only one significant result involving onset: a greater VSPE for the /b/ than the /g/ onset (respectively −64 ms vs. −47 ms when collapsed across fidelity) ($F\ (1{,}128) = 18{\cdot}17$, $MSE = 4340{\cdot}41$, $p < \cdot 0001$, partial $\eta^2 = \cdot 124$). The 95% confidence intervals shown in Figure 5 indicated a significant VSPE – the confidence interval did not contain zero – for all datapoints excepting one; namely the intact stimuli with a /g/ onset in the five-year-olds.

In short, Analysis 1 indicates that phonological priming OVERALL was significantly greater for the audiovisual than auditory mode. Visual speech produced significantly greater phonological priming in children from four to fourteen years, with all age groups showing a significant effect of visual speech for most conditions. The influence of visual speech was slightly greater for the /b/ than the /g/ onsets, but phonological priming did not show the pronounced differences that characterize identifying phonemes on direct measures of speechreading (see also Jordan & Bevan, 1997). Next, we investigated the effect of child factors on performance as a function of the mode and stimulus fidelity.

ANALYSIS 2

To identify the child factors underpinning the VSPE, we analyzed results for the intact vs. non-intact words and nonwords as a function of the children's ages and verbal abilities. Our goal was to determine which of the child factors – among age, vocabulary, phonological awareness, and speechreading (visual only speech recognition) – uniquely contributed to performance. We defined 'uniquely' statistically as the independent contribution of each variable after controlling for the other variables (Abdi, Edelman, Valentin & Dowling, 2009). Use of this regression analytic approach, which yields part (aka, semi-partial) correlations, is essential for identifying the critical individual factors underpinning speech perception by children.

We investigated two basic research questions: Is the VSPE supported by the same unique child factors for (i) intact vs. non-intact stimuli and (ii) words vs. nonwords? There is little to no evidence to assist in predicting these results. However, we can predict the effects of child factors from models of the picture–word task. As noted in the 'Introduction', the model of Levelt *et al.* (1991) based on auditory distractors proposes that the phonologically related distractor (e.g. [picture]–[distractor] pair of [bug]–[bus]) primes picture naming by creating crosstalk between the input and output phonological representations supporting speech perception and production. The congruent distractor activates input phonological representations whose activation spreads to activate the corresponding output phonological representations, and this crosstalk speeds selection of the output speech segments for naming (Roelofs, 1997). These models – to the extent they generalize – predict that the quality of children's phonological representations or knowledge will influence performance on our task. Again, we view visual speech as an extra phonetic resource as proposed by Campbell (1988). Finally, based on the hierarchical model of speech segmentation (Mattys *et al.*, 2005), we previously proposed that children's sensitivity to visual speech will vary depending on their weighting of the phonetic–phonological content. If this is so, the children's phonological knowledge may be uniquely important to the VSPE, particularly for nonwords. In short, the findings below should provide fundamental new knowledge about the contribution of age-related improvements vs. the absolute excellence of selected verbal skills to speech perception by children.

METHODS

Participants

Participants were the four groups of Analysis 1.

Materials and procedure

Receptive vocabulary was estimated with the Peabody Picture Vocabulary Test (Fourth Edition; Dunn & Dunn, 2007), measuring children's ability to identify a picture illustrating a spoken word's meaning. Phonological awareness was estimated with three subtests of the Pre-Reading Inventory of Phonological Awareness (Dodd, Crosbie, McIntosh, Teitzel & Ozanne, 2003), measuring children's ability to isolate onset phonemes, recognize alliterative onset phonemes, and segment the phonemes within a word. Speechreading was estimated with the Children's Audio-Visual Enhancement Test (Tye-Murray & Geers, 2001), measuring children's ability to repeat words presented in the visual (and auditory) modes. Results for the auditory mode were not reported because all age groups performed at ceiling. Results for the visual mode were scored by words and by word onsets with visemes (visually indistinguishable phonemes) counted as correct. The latter results were used to quantify speechreading for the regression analyses.

RESULTS

Descriptive statistics for child factors

Table 3 summarizes the average ages along with selected verbal skills in the groups. Vocabulary knowledge in the groups averaged about 120 standard score, a result indicating that these children had higher than average verbal skills. Although high verbal performance is, in general, typical of children in research studies, such performance could potentially affect the generalizability of the results to children with more 'average' verbal abilities. Phonological awareness averaged 58% correct in the youngest group and about 81% correct in the other groups; performance ranged from the ceiling in all groups to a floor of about 5% in the five-year-olds, 45% in the seven-year-olds and nine-year-olds, and 60% in the twelve-year-olds. Speechreading ranged, on average, from 6% to 25%

TABLE 3. *Averages ages and vocabulary, phonology, and speechreading in the four age groups (N = 132)*

	Age groups (yrs)			
Measures	5 $N = 38$	7 $N = 32$	9 $N = 32$	12 $N = 30$
Age (years; months)	4;11 (0·53)	7;00 (0·59)	9;02 (0·87)	12;04 (1·24)
Receptive vocabulary (standard score)	120·86 (9·70)	117·44 (11·95)	121·10 (13·57)	122·12 (10·87)
Phonological awareness (% correct)	58·22 (17·42)	80·45 (8·92)	80·35 (7·04)	83·27 (6·04)
Speechreading (percent correct) scored by words	5·80 (9·76)	10·67 (8·44)	15·30 (13·39)	25·32 (10·64)
scored by word onsets*	39·23 (20·67)	54·56 (17·71)	64·45 (15·13)	74·20 (11·65)

NOTE: standard deviations are in parentheses; * onsets were scored with visemes counted as correct (e.g. *p̲at* for *b̲at*). Each age group represents a range of chronological ages (see text).

across groups when scored by words and 39% to 74% when scored by word onsets.

Association between VSPE and child factors

The goal of this project was explanatory – thus we focused on understanding which of the child factors, if any, contributed significantly to the VSPE when the effects of the other factors were controlled. To assess the relative importance of each factor in determining the VSPE, we conducted four regression analyses ((i) words–intact, (ii) words–non-intact, (iii) nonwords–intact, and (iv) nonwords–non-intact) to obtain the part (aka semi-partial) correlation coefficients and partial F statistics (Abdi *et al.*, 2009). The dependent variable was always the VSPE, and the independent variables were always the standardized scores for age, vocabulary, phonological awareness, and speechreading.[2] Table 4 summarizes these regression results, along with the slope coefficients, for the intact vs. non-intact conditions (left vs. right panels) of the words vs. nonwords (top vs. bottom panels).

Results for the part correlations reflected one overall pattern for the intact stimuli and the non-intact words: the VSPE was uniquely influenced by the children's phonological skills. In contrast to this pattern of results, the VSPE for the low-fidelity (non-intact) nonwords was uniquely influenced only by

[2] The intercorrelations among this set of predictor variables were as follows: (i) age vs. vocabulary (0·109), phonological awareness (0·583), and visual speechreading (0·589); (ii) vocabulary vs. phonological awareness (0·106) and visual speechreading (–0·071); and (iii) phonological awareness vs .visual speechreading (0·356).

TABLE 4. *Summary of statistical results for relation between VSPE and individual child factors*

Variables	High fidelity (intact)				Low fidelity (non-intact)			
	Slope	**Part *r***	***Partial F***	***p***	**Slope**	**Part *r***	***Partial F***	***p***
				Words				
Age	0·502	·032	0·01	ns	7·805	·084	1·07	ns
Vocabulary	−3·408	·063	0·53	ns	−1·210	·000	0·49	ns
Phonology	**12·578**	**·184***	**4·44**	**·037**	**18·744**	**·235***	**7·94**	**·005**
Speechreading	−6·167	·095	1·18	ns	−1·720	·000	0·07	ns
				Nonwords				
Age	12·098	·100	1·40	ns	−8·509	·063	0·46	ns
Vocabulary	−10·690	·127	2·11	ns	−7·650	·078	0·73	ns
Phonology	**−23·246**	**·217***	**6·34**	**·013**	−0·132	·000	0·00	ns
Speechreading	−5·177	·055	0·34	ns	**26·616**	**·210***	**5·91**	**·016**

NOTES: ns = not significant (p > ·05). The part correlation coefficients and the partial F statistics evaluate the variation in VSPE uniquely accounted for (after removing the influence of the other variables) by age, vocabulary, phonology, or speechreading of onsets. The slope coefficients quantify the slope of the relationship between the VSPE and each individual child factor when all of the other child factors are held constant. The multiple correlation coefficients for all of the variables considered simultaneously were as follows: words: ·223 (intact) and ·358 (non-intact); nonwords: ·261 (intact) and ·247 (non-intact). *df*s = 1,127 for partial F and 4,127 for Multiple R.

speechreading skills. In short, these results indicate that the VSPE is underpinned by phonological skills unless the input is an unfamiliar low-fidelity stimulus without a lexical representation, in which case speechreading skills become uniquely contributory.

DISCUSSION

This research assessed the influence of visual speech on phonological priming by high- vs. low-fidelity auditory speech in children between four and fourteen years. The low-fidelity stimuli were words and nonwords with a visual consonant + rhyme coupled to an auditory non-intact onset + rhyme. Our research paradigm presented the stimuli in the auditory and audiovisual modes to determine whether (i) the presence of visual speech would fill in the non-intact auditory onsets and prime picture naming more effectively than auditory speech alone and (ii) phonological priming would display a greater influence of visual speech for non-intact than intact auditory onsets. The results showed a significant VSPE not only for the non-intact, but also for the intact, onsets – a pattern indicating that visual speech not only filled in the non-intact auditory cues but also supplemented the intact auditory cues. We observed a consistently significant influence of visual speech on phonological priming for children of all ages between four to fourteen years for most conditions. The significant boost by visual speech was substantial, particularly for the non-intact stimuli: about 34 ms (intact) and 77 ms (non-intact).

Results assessing lexical status indicated that the nonwords reflected significantly greater priming OVERALL than the words (respectively −153 ms vs. −112 ms). However, the lexical status of stimuli interacted with the mode and fidelity. Results showed that the VSPE for non-intact onsets was significantly greater for nonwords than words (respectively 91 ms vs. 62 ms), whereas the VSPE for intact onsets did not differ significantly for the nonwords vs. words (respectively 36 ms vs. 33 ms; Figure 3 collapsed across age groups). A greater VSPE for the non-intact nonwords than words is consistent with our predictions. When auditory speech has low fidelity, visual speech assumes a relatively greater weight and thus affects performance more. When this relatively greater weighting of visual speech is coupled with the relatively greater weighting of the phonetic–phonological content for nonwords, a significantly greater influence of visual speech is observed for nonwords than words.

With regard to the higher-order interaction – the VSPE differed for non-intact, but not for intact, words vs. nonwords – we should note that our set of onsets was constrained (word or nonword stimuli consisting of /b/ and /g/ onsets along with filler and baseline items). Thus, it is possible that all of the INTACT word/nonword onsets in this limited set had sufficient sensory input for correct perception, and this would yield no difference in performance for the intact words vs. nonwords.

Results for the multiple comparisons – in all age groups – indicated significantly greater priming for the audiovisual than the auditory mode not only for all non-intact but also for all intact conditions excepting intact nonwords in the five-year-olds. A worthy question is: Why did these results – in contrast to the literature – show a significant VSPE for intact stimuli in all age groups? One possibility is that the variability introduced by intermixing the fidelity (intact vs. non-intact) and mode (audiovisual vs. auditory) of the stimuli may have increased children's awareness of the sensory qualities of the input – thus making visual speech more potent. Results on the Garner task clearly indicate that participants – when they classify consonants – find it harder to ignore irrelevant inputs that vary (/ba/, /bi/ vs. /ga/, /gi/) vs. those that are constant (/ba/ vs. /ga/). This pattern suggests that the children may have found it harder to ignore speech distractors that varied in both fidelity and mode. Results on the Garner task would appear to generalize to our task because individuals process speech automatically (even when instructed to attend to picture naming) and implicitly encode and integrally process all speech cues, not just the target cues. To illustrate, three- to five-year-olds on a talker recognition task identify cartoon characters from their vocal signatures (e.g. pitch, speaking rate, dialect) at well above chance levels, indicating that these non-target speech cues were incidentally learned (Spence, Rollins & Jerger, 2002). With regard to age, Jerger and colleagues (1993) have assessed performance on the Garner task with other types of speech cues and observed integral processing at all ages between three and seventy-nine years. Thus, we propose that the variability in both stimulus fidelity and mode may have made visual speech more effective at influencing performance. This reasoning is consistent with the proposals of dynamic systems theory (see 'Introduction'; Smith & Thelen, 2003).

Another relevant question concerned whether the non-intact audiovisual distractors were as phonologically effective as their intact counterparts (in other words, was the Visual Speech Fill-In Effect complete or partial?). Results in all age groups indicated that the intact audiovisual distractors produced greater phonological priming than their non-intact counterparts. Thus, even though the Visual Speech Fill-In Effect for non-intact distractors was impressively robust, the non-intact audiovisual distractors were not as phonologically potent as their intact counterparts. This outcome agrees with previous results indicating that the visually influenced percept of the McGurk effect is not equivalent to the percept produced by a comparable audiovisual syllable (Rosenblum & Saldana, 1992).

Finally, results assessing the child factors underpinning performance indicated that the VSPE was uniquely influenced by phonological skills for the intact words and nonwords and the non-intact words. In contrast to this unified pattern of results, the VSPE for non-intact nonwords was uniquely influenced by speechreading skills. We can speculate that the influence of visual speech is more data-driven – i.e., more dependent on speechreading the 'data' – when the input is unfamiliar non-intact nonwords, and more knowledge-driven – i.e., more dependent on phonological skills – when the input is intact words/nonwords or familiar non-intact words with stored lexical phonological patterns. Clearly the factors associated with the influence of visual speech on performance are multi-faceted.

In conclusion, the new Visual Speech Fill-In Effect extends the range of measures for assessing benefit from visual speech by children. Results on the new measure document that children from four to fourteen years benefit from visual speech during multimodal speech perception. These findings emphasize that children – like adults – experience a speaker's multimodal utterance. Such information seems critical for incorporating visual speech into our developmental theories of speech perception.

SUPPLEMENTARY MATERIALS

For supplementary material for this paper, please visit http://dx.doi.org/10.1017/S030500091500077X.

REFERENCES

Abdi, H., Edelman, B., Valentin, D. & Dowling, W. (2009). *Experimental design and analysis for psychology*. New York: Oxford University Press.

Abdi, H. & Williams, L. (2010). Contrast analysis. In N. Salkind (ed.), *Encyclopedia of research design*, 243–51. Thousand Oaks, CA: Sage.

Barutchu, A., Crewther, S., Kiely, P., Murphy, M. & Crewther, D. (2008). When /b/ill with /g/ill becomes /d/ill: evidence for a lexical effect in audiovisual speech perception. *European Journal of Cognitive Psychology* **20**, 1–11.

Beery, K. & Beery, N. (2004). *The Beery-Buktenica Developmental Test of Visual-Motor Integration with Supplemental Developmental Tests of Visual Perception and Motor Coordination*, 5th ed. Minneapolis: NCS Pearson, Inc.

Bell-Berti, F. & Harris, K. (1981). A temporal model of speech production. *Phonetica* **38**, 9–20.

Bonte, M. & Blomert, L. (2004). Developmental changes in ERP correlates of spoken word recognition during early school years: a phonological priming study. *Clinical Neurophysiology* **115**, 409–23.

Brancazio, L. (2004). Lexical influences in audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance* **30**, 445–63.

Brooks, P. & MacWhinney, B. (2000). Phonological priming in children's picture naming. *Journal of Child Language* **27**, 335–66.

Calvert, G., Campbell, R. & Brammer, M. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology* **10**, 649–57.

Campbell, R. (1988). Tracing lip movements: making speech visible. *Visible Language* **22**, 32–57.

Damian, M. & Dumay, N. (2007). Time pressure and phonological advance planning in spoken production. *Journal of Memory and Language* **57**, 195–209.

Damian, M. & Martin, R. (1999). Semantic and phonological codes interact in single word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **25**, 345–61.

Dekle, D., Fowler, C. & Funnell, M. (1992). Audiovisual integration in perception of real words. *Perception & Psychophysics* **51**, 355–62.

Desjardins, R., Rogers, J. & Werker, J. (1997). An exploration of why preschoolers perform differently than do adults in audiovisual speech perception tasks. *Journal of Experimental Child Psychology* **66**, 85–110.

Dodd, B. (1977). The role of vision in the perception of speech. *Perception* **6**, 31–40.

Dodd, B., Crosbie, S., McIntosh, B., Teitzel, T. & Ozanne, A. (2003). *Pre-Reading Inventory of Phonological Awareness*. San Antonio, TX: Psychological Corporation.

Dunn, L. & Dunn, D. (2007). *The Peabody Picture Vocabulary Test-IV*, 4th ed. Minneapolis, MN: NCS Pearson.

Easton, R. & Basala, M. (1982). Perceptual dominance during lipreading. *Perception & Psychophysics* **32**, 562–70.

Erdener, D. & Burnham, D. (2013). The relationship between auditory-visual speech perception and language-specific speech perception at the onset of reading instruction in English-speaking children. *Journal of Experimental Child Psychology* **114**, 120–38.

Evans, J. (2002). Variability in comprehension strategy use in children with SLI: a dynamical systems account. *International Journal of Language and Communication Disorders* **37**, 95–116.

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**, 457–67.

Garner, W. (1974). *The processing of information and structure*. Potomax, MD: Erlbaum.

Goldman, R. & Fristoe, M. (2000). *Goldman Fristoe 2 Test of Articulation*. Circle Pines, MN: American Guidance Service.

Gow, D., Melvold, J. & Manuel, S. (1996). How word onsets drive lexical access and segmentation: evidence from acoustics, phonology, and processing. *Spoken Language ICSLP Proceedings of the 4th International Conference* **1**, 66–9.

Grant, K., van Wassenhove, V. & Poeppel, D. (2004). Detection of auditory (cross-spectral) and auditory-visual (cross-modal) synchrony. *Speech Communication* **44**, 43–53.

Green, K. (1998). The use of auditory and visual information during phonetic processing: implications for theories of speech perception. In R. Campbell, B. Dodd & D. Burnham (eds), *Hearing by eye II: advances in the psychology of speechreading and auditory-visual speech*, 3–25. Hove: Taylor & Francis.

Green, K. & Kuhl, P. (1989). The role of visual information in the processing of place and manner features in speech perception. *Perception & Psychophysics* **45**, 34–42.

Hillock, A. R., Powers, A. R. & Wallace, M. T. (2011). Binding of sights and sounds: age-related changes in multisensory temporal processing. *Neuropsychologia* **49**, 461–7.

Holt, R. F., Kirk, K. I. & Hay-McCutcheon, M. (2011). Assessing multimodal spoken word-in-sentence recognition in children with normal hearing and children with cochlear implants. *Journal of Speech, Language, and Hearing Research* **54**, 632–57.

Huyse, A., Berthommier, F. & Leybaert, J. (2013). Degradation of labial information modifies audiovisual speech perception in cochlear-implanted children. *Ear and Hearing* **34**, 110–21.

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 multimodal picture-word task. *Journal of Experimental Child Psychology* **102**(1), 40–59.

Jerger, S., Damian, M. F., Tye-Murray, N. & Abdi, H. (2014). Children use visual speech to compensate for non-intact auditory speech. *Journal of Experimental Child Psychology* **126**, 295–312.

Jerger, S., Martin, R. & Damian, M. F. (2002). Semantic and phonological influences on picture naming by children and teenagers. *Journal of Memory and Language* **47**, 229–49.

Jerger, S., Pirozzolo, F., Jerger, J., Elizondo, R., Desai, S., Wright, E. & Reynosa, R. (1993). Developmental trends in the interaction between auditory and linguistic processing. *Perception & Psychophysics* **54**, 310–20.

Jordan, T. & Bevan, K. (1997). Seeing and hearing rotated faces: influences of facial orientation on visual and audiovisual speech recognition. *Journal of Experimental Psychology: Human Perception and Performance* **23**, 388–403.

Lalonde, K. & Holt, R. (2015). Preschoolers benefit from visually salient speech cues. *Journal of Speech, Language, and Hearing Research* **58**, 135–50.

Levelt, W., Schriefers, H., Vorberg, D., Meyer, A., Pechmann, T. & Havinga, J. (1991). The time course of lexical access in speech production: a study of picture naming. *Psychological Review* **98**, 122–42.

Locke, J. (1993). *The child's path to spoken language*. Cambridge, MA: Harvard University Press.

Mattys, S. (2014). Speech perception. In D. Reisberg (ed.), *The Oxford handbook of cognitive psychology*, 391–411. Oxford: Oxford University Press.

Mattys, S. L., White, L. & Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: a hierarchical framework. *Journal of Experimental Psychology-General* **134**, 477–500.

McGurk, H. & McDonald, M. (1976). Hearing lips and seeing voices. *Nature* **264**, 746–8.

Mills, A. (1987). The development of phonology in the blind child. In B. Dodd & R. Campbell (eds), *Hearing by eye: the psychology of lipreading*, 145–61. London: Erlbaum.

Nath, A. & Beauchamp, M. (2011). Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *Journal of Neuroscience* **31**, 1704–14.

Roelofs, A. (1997). The WEAVER model of word-form encoding in speech production. *Cognition* **64**, 249–84.

Rosenblum, L. & Saldana, H. (1992). Discrimination tests of visually influenced syllables. *Perception & Psychophysics* **52**, 461–73.

Ross, L., 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**, 2329–37.

Ross, M. & Lerman, J. (1971). *Word Intelligibility by Picture Identification*. Pittsburgh: Stanwix House, Inc.

Sams, M., Manninen, P., Surakka, V., Helin, P. & Katto, R. (1998). McGurk effect in Finnish syllables, isolated words, and words in sentences: effects of word meaning and sentence context. *Speech Communication* **26**, 75–87.

Schriefers, H., Meyer, A. & Levelt, W. (1990). Exploring the time course of lexical access in language production: picture-word interference studies. *Journal of Memory and Language* **29**, 86–102.

Seewald, R. C., Ross, M., Giolas, T. G. & Yonovitz, A. (1985). Primary modality for speech perception in children with normal and impaired hearing. *Journal of Speech and Hearing Research* **28**, 36–46.

Sekiyama, K. & Burnham, D. (2004). Issues in the development of auditory-visual speech perception: adults, infants, and children. *Interspeech-2004*, 1137–40.

Sekiyama, K. & Burnham, D. (2008). Impact of language on development of auditory-visual speech perception. *Developmental Science* **11**, 306–20.

Sekiyama, K. & Tohkura, Y. (1991). McGurk effect in non-English listeners: few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of America* **90**, 1797–805.

Smith, L. & Thelen, E. (2003). Development as a dynamic system. *Trends in Cognitive Sciences* **7**, 343–8.

Spence, M., Rollins, P. & Jerger, S. (2002). Children's recognition of cartoon voices. *Journal of Speech, Language, and Hearing Research* **45**, 214–22.

Stevenson, R. A., Wallace, M. T. & Altieri, N. (2014). The interaction between stimulus factors and cognitive factors during multisensory integration of audiovisual speech. *Frontiers in Psychology* **5**, 352. doi: 10.3389/fpsyg.2014.00352.

Storkel, H. & Hoover, J. (2010). An on-line calculator to compute phonotactic probability and neighborhood density based on child corpora of spoken American English. *Behavior Research Methods* **42**(2), 497–506.

ten Oever, S., Sack, A. T., Wheat, K. L., Bien, N. & van Atteveldt, N. (2013). Audio-visual onset differences are used to determine syllable identity for ambiguous audio-visual stimulus pairs. *Frontiers in Psychology* **4**: 331. doi: 10.3389/fpsyg.2013.00331.

Tomiak, G., Mullennix, J. & Sawusch, J. (1987). Integral processing of phonemes: evidence for a phonetic mode of perception. *Journal of the Acoustical Society of America* **81**, 755–64.

Tremblay, C., Champoux, R., Voss, P., Bacon, B., Lepore, F. & Theoret, H. (2007). Speech and non-speech audio-visual illusions: a developmental study. *PLoS One* **2**(8), e742. doi:10.1371/journal.pone.0000742.

Tye-Murray, N. (2014). *Foundations of aural rehabilitation: children, adults, and their family members*, 4th ed. Boston: Cengage Learning.

Tye-Murray, N. & Geers, A. (2001). *Children's Audio-Visual Enhancement Test*. St Louis, MO: Central Institute for the Deaf.

Vatakis, A. & Spence, C. (2007). Crossmodal binding: evaluating the 'unity assumption' using audiovisual speech stimuli. *Perception & Psychophysics* **69**(5), 744–56.

Vitevitch, M. & Luce, P. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments & Computers* **36**, 481–7.

Wightman, F., Kistler, D. & Brungart, D. (2006). Informational masking of speech in children: auditory-visual integration. *Journal of the Acoustical Society of America* **119**, 3940–9.