Language–General Auditory–Visual Speech Perception: Thai–English and Japanese–English McGurk Effects

Denis Burnham^{1,*} and Barbara Dodd²

¹ MARCS Institute for Brain, Behaviour and Development, Western Sydney University,
Sydney, Australia

² University of Queensland, Brisbane, Australia

Received 13 January 2017; accepted 19 June 2017

Abstract

Cross-language McGurk Effects are used to investigate the locus of auditory–visual speech integration. Experiment 1 uses the fact that [ŋ], as in ‘sing’, is phonotactically legal in word-final position in English and Thai, but in word-initial position only in Thai. English and Thai language participants were tested for ‘n’ perception from auditory [m]/visual [ŋ] (A[m]V[ŋ]) in word-initial and -final positions. Despite English speakers’ native language bias to label word-initial [ŋ] as ‘n’, the incidence of ‘n’ percepts to A[m]V[ŋ] was equivalent for English and Thai speakers in final *and* initial positions. Experiment 2 used the facts that (i) [ð] as in ‘that’ is not present in Japanese, and (ii) English speakers respond more often with ‘tha’ than ‘da’ to A[ba]V[ga], but more often with ‘di’ than ‘thi’ to A[bi]V[gi]. English and three groups of Japanese language participants (Beginner, Intermediate, Advanced English knowledge) were presented with A[ba]V[ga] and A[bi]V[gi] by an English (Experiment 2a) or a Japanese (Experiment 2b) speaker. Despite Japanese participants’ native language bias to perceive ‘d’ more often than ‘th’, the four groups showed a similar phonetic level effect of [a]/[i] vowel context × ‘th’ vs. ‘d’ responses to A[b]V[g] presentations. In Experiment 2b this phonetic level interaction held, but was more one-sided as very few ‘th’ responses were evident, even in Australian English participants. Results are discussed in terms of a phonetic plus postcategorical model, in which incoming auditory and visual information is integrated at a phonetic level, after which there are post-categorical phonemic influences.

Keywords

Auditory–visual integration, cross-language speech perception, phonetic and phonemic speech perception, the McGurk effect

* To whom correspondence should be addressed. E-mail: denis.burnham@westernsydney.edu.au

1. Introduction

Visual (face and lip) information plays an important role in speech perception, not only in noisy but otherwise natural conditions (Sumby and Pollack, 1954), but also in clear but unnatural conditions when auditory and visual speech components are mismatched (Dodd, 1977; McGurk and MacDonald, 1976). These pioneering mismatching studies plus four decades of research on what is now called the McGurk Effect show that visual speech information is combined with auditory speech information, albeit unconsciously, whenever it is available (although the nature of the combination may differ from more natural conditions; see Alsius *et al.* in this issue). The original McGurk version — auditory [ba] dubbed onto visual [ga] (A[ba]V[ga]) — is particularly compelling because the result is an emergent percept, ‘da’ or ‘tha’, thus possibly implicating auditory–visual integration. In some other auditory–visual mismatches there is, rather than integration, more simple modality dominance, as in the case of A[ba]V[va] being perceived as ‘va’; or modality combination, such as A[ga]V[ba] being perceived as ‘bga’. Such combinations seems to occur in A[velar]V[bilabial], not A[bilabial]V[velar] presentations; in the McGurk and MacDonald (1976) study there were no ‘combination’ responses (e.g., *baga*, *bagba*) for A[ba]V[ga] or A[pa]V[ka], which generally yielded ‘da’ or ‘ta’ responses, respectively, but when the auditory component was a velar rather than a bilabial, A[ga]V[ba] or A[ka]V[pa] there were 54% and 44% combination responses, respectively. The issue of phonotactically illegal responses such as ‘bga’ is taken up further in the introduction to Experiment 1.

These early demonstrations sparked investigation of the issues and processes involved in the McGurk effect and auditory–visual integration (see e.g., Bernstein *et al.*, 2002; Robert-Ribes *et al.*, 1995, 1996). These issues include whether auditory–visual integration is specific to speech; where, how, and when the McGurk effect occurs; and the McGurk effect within and between languages. Research in each of these areas is reviewed, especially with respect to how they relate to the psycholinguistic locus of auditory–visual speech integration, which is the focus of the two studies presented here.

1.1. Auditory–Visual Integration: Speech vs. Non-Speech

McGurk-type effects have been found with non-speech stimuli, e.g., when auditory and visual information for ‘pluck’ and ‘bow’ sounds on a cello are mismatched (Saldaña and Rosenblum, 1993). In a second experiment Saldaña and Rosenblum found the music effect to be weaker than a consonant mismatch McGurk effect, suggesting that, while the McGurk effect may not be specific to speech, different processes may be involved in speech and non-speech McGurk-like effects. This is supported by neural evidence (Baart *et al.*, 2014) that, while peak N1 EEG responses occur earlier in auditory–visual than

auditory-alone conditions in both speech and non-speech (sine-wave-speech) modes, only in the speech mode is there modulation of the P2 response and an effect of phonetically-incongruent auditory–visual stimuli on event-related potentials (ERPs) around 200 ms after stimulus onset. Moreover, there is fMRI evidence that lipreading activates the auditory cortex, but only while meaningful speech movements are observed, not for non-speech (gurning) movements (Calvert *et al.*, 1997; Campbell *et al.*, 2001). Thus auditory–visual and visual processing differs for speech and non-speech, though the exact nature of this difference is yet to be specified.

1.2. Auditory–Visual Integration: Where?

Based on evidence that auditory–visual integration in primates occurs in the superior temporal sulcus (STS), Beauchamp, Nath and Pasalar (2010) investigated the role of the STS in the McGurk effect. They found the incidence of adults' fusion responses to A[ba]V[ga] (but not auditory-only [ba]) was significantly reduced when a transcranial magnetic stimulation (TMS) signal was delivered to the STS within 100 ms of auditory stimulus onset. This temporal constraint is consistent with other findings: ERPs for incongruent auditory–visual speech presentations 200 ms after stimulus onset (Baart *et al.*, 2014); speeding up of cortical processing of auditory signals if visual speech is presented within 100 ms of auditory signal onset (Van Wassenhove *et al.*, 2015); and fusion responses to McGurk effect stimuli even up to an auditory–visual temporal asynchrony window of 200 ms (Munhall *et al.*, 1996; Van Wassenhove *et al.*, 2007).

1.3. Auditory–Visual Integration: how?

The issue of how auditory–visual occurs is addressed below, first in a non-linguistic and then in a linguistic manner.

1.3.1. Common or Separate Metrics for Auditory and Visual Information

A broad distinction between the possible processes involved in the McGurk effect is between integration of auditory and visual information in a common metric space *vs.* the combination of modality-specific auditory and visual information in a sequential decision process.

The sequential process view is well exemplified by Massaro's (1987, 1996) fuzzy logic model of speech perception (FLMP). In this, incoming auditory, visual, and top-down information are processed independently and in parallel, then evaluated with features in each source being independently assigned fuzzy truth values for the degree to which they match phoneme prototype features, and finally there is specification of a particular phoneme. Thus auditory and visual perceptual information remain separate rather than being represented in a common metric.

In an opposing view information originating from auditory and visual channels are represented in a common metric, which has been proposed to be auditory (Summerfield, 1987), gestural or articulatory (Liberman and Mattingly, 1985, 1989; McGurk and Buchanan, 1981), or amodal (Kuhl and Meltzoff, 1984, 1988; Studdert-Kennedy, 1986).

1.3.2. *Phonetic or Phonemic Auditor–Visual Integration*

Another way in which this distinction may be couched is between phonetic and phonemic modes of psycholinguistic processing. The phonetic *vs.* phonemic distinction rests on language experience. A large number of speech sounds, *phones*, are used across the world's languages, but only a subset of phones or groupings of phones, *phonemes*, are used to distinguish meaning in a particular language. For example, the phones [tʃ] and [ʃ], as in 'chop' and 'shop' respectively, are separate phonemes in English but in Thai they are allophones within the same phoneme class, the 'ch' sound, ฉ. So in Thai [tʃ] *vs.* [ʃ] is a phonetic distinction and not used in everyday speech, whereas in English [tʃ] *vs.* [ʃ] is a phonemic distinction that distinguishes meaning.

Auditory–visual integration at a phonetic level would be in a space in which all language-general phonetic information was available free from specific linguistic experience; conversely auditory–visual integration at a phonemic level would allow language-specific information to impinge upon the percept, or perhaps more correctly upon the participants' verbal responses. As common metric theories emphasise the abstract nature of this metric, it would be more in line with phonetic processing. As separate metric theories rest on comparison of truth values to learned prototypes, they would be more in line with phonemic processing.

An ingenious set of experiments by Green (Green and Kuhl, 1989, 1991; Green and Miller, 1985; Green *et al.*, 1991) address this issue by examining whether established *auditory* speech perception phenomena also occur when some phonetic information is specified auditorally, and some *visually*. For example, when identifying speech sounds along an acoustic [bi]–[pi] voice onset time (VOT) continuum, English language participants have higher VOT crossovers (more [bi] responses) when the vowel length of [i] is increased. To test this cross-modally, Green and Miller (1985) paired an auditory [bi]–[pi] continuum spoken at medium rate with visual articulations of [bi] and [pi] spoken at slow and fast rates, and found that visual speech rate information acted upon auditory perception in the same way as did auditory rate information. These results show psycholinguistic level interaction of auditory and visual information, and that visemes independently specify phonetic information, just as their auditory counterparts do. However, the psycholinguistic level at which this occurs is not wholly specified. To determine this, an extension would be required in which the same auditory–visual integration is specified

in a language in which vowel length is phonemic. If the same effects were still obtained then this would support a phonetic level explanation as the change in vowel length would entail a change of a phoneme not just a phonetic change. *A similar, but not identical line of reasoning is employed in the two experiments reported here.*

1.4. Auditory–Visual Integration: When?

The common metric and phonetic views imply auditory–visual integration temporally early in psycholinguistic processing, while prototype and phonemic views imply temporally later processing. Visual speech activates auditory cortex (see Calvert *et al.*, 1997; Campbell *et al.*, 2001, see above) and brainstem FFRs (Frequency Following Responses) show that as early as 11 ms post-acoustic stimulus onset, auditory–visual speech compared with unimodal speech results in a 1.3 ms delay at the brainstem, with slightly greater suppression for congruent, A[da]V[da], than incongruent, A[da]V[fu], AV pairings (Musacchia *et al.*, 2006).

One interpretation of these studies is that auditory–visual integration and/or visual influence occurs in an auditory brainstem common metric (with differential treatment of congruent and incongruent McGurk-like stimuli, Musacchia *et al.*, 2006), and that such auditory coding continues through to the auditory cortex both generally (Calvert *et al.*, 1997; Campbell *et al.*, 2001, see above) and differentially, as fMRI measures show there is greater posterior parietal cortex activation for concordant stimuli than discordant auditory–visual stimuli (Saito *et al.*, 2005). However, further evidence is required before definitive conclusions can be drawn.

1.5. Auditory–Visual Integration: Within and Between Languages

Phoneme classes are the product of linguistic experience. At birth, infants discriminate between phones irrespective of whether they are native or non-native in the ambient speech environment. Then gradually and especially from around four months infants learn to perceive speech in terms of the phoneme classes of the surrounding (native) language or languages. The attunement process is similar across languages, but the result differs — particular language-specific phonemic classes are set up over and above the phonetic substrate. As attunement is attentional rather than structural, there is no absolute loss of perceptual discrimination and within-phoneme (phonetic) information remains available. *Thus investigation of auditory–visual integration within and between languages should prove instructive for the issue of phonetic vs. phonemic level processing.*

1.5.1. Auditory–Visual Integration: Development Within Languages

The McGurk effect is evident early in infancy (Burnham and Dodd, 2004; Desjardins and Werker, 2004; Rosenblum *et al.*, 1997). For example, 4 1/2-month-old infants perceive A[ba]V[ga] as ‘da’ or ‘tha’ significantly more often than as ‘ba’ (Burnham and Dodd, 2004). Such evidence *could* be taken to show auditory–visual integration occurs before phonemic processing is evident in development. However, as with the Green studies and their relation to phonetic processing, this is evidence not watertight, as there are indications of phonemic influences even at this tender age.

First, it has been found that English language 10-week-olds match a voice to a talking face, irrespective of whether the face–voice matching is for a native English or a foreign voice; but 20-week-olds only do so for their native English language (Dodd and Burnham, 1988). Such auditory–visual speech perception language-specificity is also evident in visual speech; English-language 4- and 6-month-olds discriminate English and French languages on the basis of visual information alone, but 8-month-olds only do so if they have bilingual English/French experience (Weikum, Vouloumanos, Navarra, Soto-Faraco, Sebastian-Galles and Werker, 2007).

Second, there is recent evidence that infants as young as three months have already set up phonological categories. Thirty-year-old adults who were adopted at three to five months from Korean families and brought up in Dutch families in the Netherlands showed, compared to 30-year-old Dutch controls, better perceptual learning of a now foreign Korean three-way voicing contrast, better generalisation of such learning to different speech (voicing) contrasts, more accurate production of the contrasts, and significant speech perception–production correlations compared to inferior perception and no such correlations in controls (Choi *et al.*, 2017). These results show that phoneme categories are set up before three to five months and are remarkably strong and resilient. Thus, while 4 1/2-month-old infants’ perception of the McGurk effect is striking, it is not necessarily achieved in the absence of phonological categories, so the occurrence of the McGurk effect in infancy is silent with respect to whether phonetic or phonemic processes are involved in the mature McGurk effect. *The studies reported here are designed to provide further evidence on the influence of phonetic and phonemic factors on McGurk effect responses.*

1.5.2. Auditory–Visual Integration: Development Between Languages

The incidence and nature of the McGurk effect changes as a function of language in three ways. First, the McGurk effect is stronger in some languages than others. In a series of studies Sekiyama and her colleagues found a much weaker McGurk effect in Japanese than in English listeners: Japanese listeners are much less influenced by visual speech (Sekiyama and Tohkura,

1991). However, Japanese listeners do increase their reliance on visual information when auditory noise is introduced (Sekiyama, 1994) and Sekiyama *et al.* (1995) found other conditions (when easy-to-lipread speakers are used, or when speakers lengthen their vowels) under which Japanese listeners' incorporation of visual cues is facilitated. Nevertheless, Japanese listeners generally show less robust McGurk effects than English language listeners. Sekiyama and Burnham (2008) discovered the developmental locus of this departure; while Japanese- and English-language six-year-olds show equivalent McGurk effect strength, between six and eight years McGurk effect strength in English-language children increases, while that for Japanese children remains unchanged between six and eight years, and even beyond, at 11 years and in adulthood. This six- to eight-year-old increase for English language children is thought to be due to the relative complexity, compared to Japanese, of the English phonological repertoire and its more complex phonotactics, combined with the opacity of the English script (which is being learned between six and eight years) (Erdener and Burnham, 2013; Sekiyama and Burnham, 2008).

Second, there are inter-language effects on auditory–visual speech integration. In a 2×2 design Sekiyama and Tohkura (1993) investigated American and Japanese listeners' perception of the McGurk effect with stimuli spoken by American and Japanese speakers. Over and above Japanese listeners' generally reduced visual influence, both Japanese and American listeners demonstrated greater visual influence when presented with stimuli in the non-native language. Sekiyama and Tohkura (1993) explained this in terms of auditory ambiguity: when participants hear phones that are phonemically relevant in their language but acoustically deviant, then any extra information that can be used, will be used. Similar foreign speaker effects have been found both for Japanese and American speakers (Kuhl, Tsuzaki, Tohkura and Meltzoff, 1994), and also in a range of other cross-language contexts: When Austrian and Hungarian listeners are presented with an Austrian speaker McGurk effect stimulus, more McGurks (more visual influence) are found in the Hungarian listeners; deGelder, Bertelson, Vroomen and Chen (1995) found McGurk fusion effects in both Dutch and Cantonese language participants listening to a Dutch speaker, but the incidence of visually-influenced blends was greater for the Cantonese listeners; and Fuster-Duran (1996) found that when presented with auditory–visual conflict in pairs of German words, e.g., 'brat–Grad', or in Spanish words e.g., 'napa–paca', Spanish and German participants incorporated visual information more for foreign than for native auditory–visual words.

Third, in a cross-language study Werker *et al.* (1992) paired auditory [ba] with visual [ba], [va], [ða], [da], [ʒa], and [ga] and asked English language speakers and French language participants with varying levels of English

language proficiency for perceptual identifications. All these consonants are phonemic in both languages, except [ð] which is not phonemic in French. There were more ‘tha’ identifications of A[ba]V[ða] by English than by French participants; the latter tended to respond with ‘da’ or ‘ta’. However, the frequency of ‘tha’ identifications increased as a function of English language experience. This linguistic influence could be perceptual: French participants may, through their language experience, have a phonologically-determined *perceptual bias* against perceiving [ð], which is ameliorated by English-language experience. However, it could equally be a post-perceptual labelling effect: French participants may *perceive* [ð] just as often as their English-speaking counterparts but could have a phonologically-determined *response bias* against reporting their perceptual experience of [ð] as ‘th’. While Werker *et al.* (1992) ensured that their French participants could produce [ð] and also found no difference in French participants’ incidence of written and spoken ‘d’ or ‘t’ responses to acoustic-only [ð] presentations, it is nevertheless possible that there was a top-down language-based response bias against a phone not used in the perceivers’ native phonology or orthography. *The experiments reported here are aimed at obviating such ambiguities, by setting up McGurk effect situations in which both phonetic and phonological information are available and attempting to specify which in fact is used.*

This review shows that the McGurk effect may be viewed in two ways. On the one hand the McGurk effect and auditory–visual integration may occur temporally early in processing, in a common abstract auditory–visual space and involve language-general phonetic psycholinguistic processes. On the other, it may occur temporally late in processing, in a disparate auditory and visual space and involve language-specific phonemic psycholinguistic processes. The evidence provides some convergence on where the McGurk effect and auditory–visual integration may occur (e.g., Beauchamp *et al.*, 2010), the temporally early fusion of auditory and visual information (e.g., Musacchia *et al.*, 2006), the effect of low level phonetic information (e.g., Green and Miller, 1985), the early appearance of auditory–visual integration in development (e.g., Burnham and Dodd, 2004; Dodd and Burnham, 1988), and between language differences (Sekiya and Burnham, 2008). But whether the McGurk effect is a phonetic or phonemic phenomenon cannot yet be ascertained. The review sets out the nature of this distinction and how it is tied in with other issues, early *vs.* late, common *vs.* disparate space. The experiments reported here bear directly on this issue and will provide information on the nature of the psycholinguistic influences on the McGurk effect and possibly more generally, on auditory–visual integration.

2. General Hypotheses

The two views of the nature of the McGurk effect set out above may be represented in the following two hypotheses. If:

- *Auditory–visual speech integration is language-specific, phonemic*, then the prevalence of the McGurk effect should be sensitive to the phoneme classes of the native language of the perceiver; the same set of McGurk stimuli (e.g., an AxVy combination) should result in different emergent percepts depending on the native language of the perceiver, and
- *Auditory–visual speech integration is language-general, phonetic*, then the same set of McGurk stimuli, AxVy, would result in the same emergent percept, irrespective of the perceiver's language background.

Two experiments designed to test these two alternatives are presented here. Each uses between-language differences in order to tease out phonetic vs. phonemic effects. In the first experiment differences in phonotactic constraints between English and Thai are the research tool, and in the second differences in phoneme repertoire between English and Japanese are the tool. In addition, both use the McGurk effect as a research tool. While it is clear from the above review that incongruent McGurk and congruent auditory–visual may differ in their integration processes, any McGurk effect results here will point the way for further auditory–visual integration research.

3. Experiment 1: Phonotactic Constraint Differences — the A[m]V[n] Fusion in Thai vs. English

The logic of this experiment rests on two facts. First, in English, the A[m]V[n] auditory–visual pairing (which incorporates visual [ŋ], as in *sing*) results in the emergent percept 'n'. Second, in Thai the use of the phoneme [ŋ] is phonotactically unconstrained, it can be used in word-initial, -medial or -final positions, whereas in English it is constrained to use in word-medial and -final positions; *it cannot be used in word-initial position*. There are mentions of phonotactically illegal phone strings in previous literature. Even the original McGurk and MacDonald (1976) paper mentions possible responses of 'bga' which, while phonotactically illegal in the written form, is probably not in the spoken form, as two spoken stop consonants must have at least a voice onset gap between them (and probably an inserted schwa sound). Here, it is not a phonotactically illegal *response* that is the target, but rather the use of [ŋ] in the A[m]V[n] stimulus in the initial position, which is of interest because of the phonotactic illegality of [ŋ] in initial position in English.

Thai and English adults were presented with A[m]V[n], either in syllable-final or syllable initial position, and with other control (auditory–visual,

auditory-only and visual-only) stimuli. Given the phonotactic illegality of initial [ŋ] in English and the ambiguity of visual-only [ŋ] (Mills, 1987; Mills and Thiem, 1980), English language participants would be expected to respond ‘n’ more often in the initial than in the final position for visual-only [ŋ] and matching auditory–visual A[ŋ]V[ŋ]. Over and above this, if auditory–visual processing occurs phonemically, then the relative incidence of ‘n’ responses to A[m]V[ŋ] should be greater in initial than final position for English but equivalent in initial than final positions for Thai participants. If, on the other hand, auditory–visual processing occurs at a more basic phonetic level, then English participants’ responses to A[m]V[ŋ] should also be equivalent in initial and final positions, and should also not differ from that of the Thai participants.

3.1. Method

3.1.1. Participants

A total of 24 native Thai and 24 native Australian English adult speakers were tested, 12 males and 12 females in each language group. The mean age of the Australian English participants was 25 years 0 months and of the Thai participants was 31 years 2 months. The Australians were all monolingual with no experience of Thai, or any language in which [ŋ] is used in word-initial position.

3.1.2. Stimulus Materials

An adult female native speaker of Central Thai with linguistic training recorded initial consonant–[a:] vowel (CV) productions — [ma:], [na:], [ŋa:], [m-ŋa], and [ŋ-ma:]; and [a:] vowel–final consonant (VC) productions — [a:m], [a:n], [a:ŋ], [a:ŋ-m], and [a:m-ŋ]. For the consonant clusters, she was asked to insert the schwa vowel, /ə/, between or after consonants, as appropriate, viz, [məŋa:], and [ŋəma:], [a:ŋmə], and [a:mŋə]. The optical and acoustic components were then separated and combined to produce auditory–visual, auditory-only, and visual-only stimuli as follows. *Auditory–visual congruent*: A[ma]V[ma], A[na]V[na], A[ŋa]V[ŋa], A[m-ŋa]V[m-ŋa], A[ŋ-ma]V[ŋ-ma]; and A[am]V[am], A[an]V[an], A[aŋ]V[aŋ], A[aŋ-m]V[aŋ-m], and A[am-ŋ]V[am-ŋ]; *Auditory–visual incongruent*: A[ma]V[ŋa], A[ŋa]V[ma]; and A[am]V[aŋ]; A[aŋ]V[am]; *Auditory-Only*: A[ma], A[ŋa]; and A[am], A[aŋ]; *Visual-Only*: V[ma], V[ŋa]; and V[am], V[aŋ]. Note that all auditory–visual stimuli were created using the same dubbing procedure. The onset of the original acoustic component of an auditory–visual stimulus was used to trigger the substitution of a new acoustic component. In visual-only conditions the original auditory component triggered the computer to play ‘silence’. This procedure was used even for the congruent auditory–visual stimuli to ensure uniformity. Each trial lasted 4 s, with 1 s of black background intervening between trials. For VO and AV trials this consisted of 1 s of a motionless face,

about 1 s of auditory–visual or visual-only articulation, and 2 s of neutral expression. For the AO trials, the speaker’s motionless face was presented for 4 s overdubbed with the appropriate speech sound.

3.2. Procedure

Participants sat in a room facing and 50 cm away from a video monitor, with a response key on a table in front of them. The response key contained a central ‘ready’ key surrounded, in a semicircle, by five response buttons, labelled ‘m’, ‘m-ng’, ‘n’, ‘ng-m’, and ‘ng’ for the English participants, and corresponding Thai graphemes for the Thai participants, ม (m), ม-ง (m-ng), น (n), ง-ม (ng-m), ง (ng).

There were two phases in the experiment, one with initial position consonant stimuli, and one with final position consonant stimuli, with phase order counter-balanced between participants, and in each phase there was a practice and a test block.

The practice block consisted of 15 practice trials; each of the five consonants/consonant clusters was presented three times, once in each of auditory–visual, auditory-only, and visual-only modes. These were [ma], [m-ŋa], [na], [ŋ-ma], and [ŋa], in the initial consonant phase, and [am], [am-ŋ], [an], [aŋ-m], and [aŋ] in the final consonant phase. These trials were included to alert participants to the possible responses they could make, to give them practice at using the five keys, and to eliminate the data of participants whose responses were too slow (see below). In practice trials an output from the experiment control program (i) activated a reward light to flash on the left side of the monitor for correct responses, and (ii) an error buzzer sound to inform the participant (and experimenter) of any failure to respond within 2.2 s, or if the participant took their finger off the ‘ready’ button prior to the onset of the sound. The participants were told the meaning of these rewards and errors. Such feedback was only given in practice trials, not in test trials.

The test blocks contained 32 trials consisting of four repetitions of the following eight stimuli: auditory-only A[m], auditory-only A[ŋ] visual-only V[m], visual-only V[ŋ], matching auditory–visual A[m]V[m], matching auditory–visual A[ŋ]V[ŋ], mismatching auditory–visual A[m]V[ŋ], and mismatching auditory–visual A[ŋ]V[m].

Instructions given at the beginning of each phase included a request to respond as quickly as possible after the stimulus had appeared on the video monitor. All trials, practice and test, proceeded as follows. To ensure concentration, any particular trial only commenced once the participant pressed the ‘ready’ key. Once the ‘ready’ key was pressed, the stimulus was presented on the video monitor and the participant’s task was to press one of the five response keys as quickly and accurately as possible. The onset of the acoustic component of the stimulus triggered a computer clock, such that participants’

button press reactions were timed from sound onset, or in the case of visual-only trials conditions from the ‘onset’ of the original but now silent auditory component.

3.3. Results

As [ŋ] is phonotactically illegal in initial but not final position in English, responses on test trials in which [ŋ] was the visual component, i.e., visual only V[ŋ], auditory–visual A[ŋ]V[ŋ], and A[m]V[ŋ] are of particular relevance here. The overall distribution of responses on such trials, the incidence of ‘n’ responses on such trials, and the reaction times for ‘n’ responses on such trials are reported in turn below.

3.4. Response Distribution on V[ŋ], A[ŋ]V[ŋ], and A[m]V[ŋ] Test Trials

Confusion matrices for each of the three stimulus types containing visual [ŋ], crossed with the five possible responses are set out in Table 1.

As can be seen in Table 1 the patterns, though not the absolute percentages of responses, are similar across the four (two participant language × word-initial/-final) conditions. Moving from V[ŋ] to A[ŋ]V[ŋ] there is invariably an increase in ‘ng’ responses and a decrease in ‘n’ responses, presumably due to the additional information for [ŋ] from the auditory modality. Moving from A[ŋ]V[ŋ] to A[m]V[ŋ], a change in the auditory component from auditory [ŋ]

Table 1.
Distribution of responses for stimuli with a Visual [ŋ] component in word-initial and word-final for Thai and Australian English participants

Response	Word-initial stimulus			Word-final stimulus		
	Vŋ	AŋVŋ	AmVŋ	Vŋ	AŋVŋ	AmVŋ
Thai participants						
m	0	0	56.25	4.18	1.05	27.08
mŋ	0	1.05	4.18	1.05	0	3.13
n	12.5	0	28.13	35.43	2.08	58.33
ŋm	0	1.05	1.05	0	0	1.05
ŋ	86.45	97.93	5.2	59.38	93.75	9.375
Australian English participants						
m	3.75	0	51.25	7.5	0	13.75
mŋ	0	1.25	1.25	0	2.5	0
n	72.5	13.75	27.5	53.75	8.75	56.25
ŋm	0	7.5	2.5	0	2.5	2.5
ŋ	22.5	76.25	15	36.25	86.25	23.75

to an auditory [m] invariably leads to a decrease in ‘ng’ responses, and an increase in two other responses: both ‘m’, a response corresponding to the auditory component of the stimulus, and ‘n’, which is not contained in either the auditory or visual component of the stimulus and what can be called an emergent fusion response. This pattern occurs in both word-initial and word-final conditions, and for both Thai and Australian English participants.

3.5. Response Incidence of ‘n’ to V[ŋ], A[ŋ]V[ŋ], and A[m]V[ŋ]

The main focus here is on the incidence of ‘n’ (fusion) responses in the three stimulus types, V[ŋ], A[ŋ]V[ŋ], and A[m]V[ŋ] in both initial and final conditions. Mean percentages of these for each of the three stimuli of interest are shown in Fig. 1. Planned contrasts within an English/Thai $\times$ (initial/final $\times$ Stimulus ([V[ŋ]/A[ŋ]V[ŋ]/A[m]V[ŋ])) analysis of variance were conducted.

3.5.1. Native Phoneme Bias

Native (with reference to the English language participants) Phoneme Bias was measured by percent ‘n’ responses to the A[ŋ]V[ŋ] stimulus. As can be seen in Fig. 1, the number of ‘n’ responses to A[ŋ]V[ŋ] was generally small but was greater for English than Thai language perceivers in English initial than English final position. There was significantly greater Native Phoneme Bias for English ($M = 9.04\%$) than for Thai perceivers (for whom the bias to ‘n’ responses to A[ŋ]V[ŋ] is *not* native) ($M = 1.05\%$), $F(1, 46) = 4.30$, and a significant English/Thai $\times$ initial/final interaction, $F(1, 46) = 6.24$. These indicate, as expected, that the most bias to perceive [n] instead of the

Figure 1. Mean percent ‘n’ responses to visual-only V[ŋ], auditory–visual A[ŋ]V[ŋ], and A[m]V[ŋ] in syllable-initial and syllable-final positions, for English and Thai perceivers. Means and standard errors.

Table 2.
Visual ambiguity scores for Thai and Australian English participants with initial and final presentations

Participant and position condition	Visual ambiguity + Phoneme bias	Phoneme bias only	Visual ambiguity
	V[ŋ]	A[ŋ]V[ŋ]	(V[ŋ])–(A[ŋ]V[ŋ])
Aus. Eng. initial	69.6	8.5	61.1
Aus. Eng. final	58.7	1.1	57.6
Thai initial	7.5	0	7.5
Thai final	29.5	2.2	27.3

auditory–visually specified [ŋ] was for English language participants in the initial position ($M = 8.5\%$).

3.5.2. Visual Ambiguity

Percentage ‘n’ responses to A[ŋ]V[ŋ] measures native phoneme bias alone, whereas percentage ‘n’ responses to V[ŋ] measures both native phoneme bias and visual ambiguity. So Visual Ambiguity alone is given by percent ‘n’ responses to the V[ŋ] stimulus minus percent ‘n’ responses to the A[ŋ]V[ŋ] stimulus. These values are presented in Table 2. There was an overall effect of visual ambiguity, $F(1, 46) = 166.91$, and an interaction of this with language group showing that visual ambiguity resulted in significantly more ‘n’ responses for English ($M = 59.35\%$) than Thai ($M = 17.4\%$) perceivers, $F(1, 46) = 30.49$. Ambiguity was also greater for final ($X = 42.5\%$) than initial consonants ($X = 34.3\%$), $F(1, 46) = 5.64$, although the English perceivers’ initial consonants condition showed the greatest visual ambiguity ($X = 61.1\%$), $F(1, 46) = 10.42$, presumably due to their inexperience in discriminating initial [n] from initial [ŋ].

3.5.3. Integration Responses

When presented with [ŋ] in syllable-initial position, English language perceivers show greater Native Phoneme Bias and greater Visual Ambiguity than do their Thai language counterparts. This is understandable given the phonotactic illegality of word-initial [ŋ] in English. If auditory–visual integration occurs at a phonemic level then the effect of this phonotactic illegality should carry over selectively to integration responses by English language participants for A[m]V[ŋ] in the initial position.

This was not the case. Analysis of ‘n’ responses to A[m]V[ŋ] revealed only one significant effect, a greater percentage of ‘n’ responses on final ($M = 59.56\%$) than initial ($M = 24.50\%$) consonants, $F(1, 46) = 34.39$.

Most importantly, there was *no* difference between English and Thai participants, nor any English/Thai $\times$ initial/final interaction. Thus, despite the significantly greater phonological bias to ‘n’ responses for initial A[ŋ]V[ŋ] and visual ambiguity for initial V[ŋ] by English speakers, their frequency of ‘n’ responses to V[ŋ] was not significantly different from that of the Thai speakers.

This is *not* due to a differential distribution of responses other than the ‘n’ (emergent fusion) responses. As can be seen in Table 1, comparison of responses across the four language $\times$ word position conditions shows that a change of the auditory component from [ŋ] to [m], i.e., from the A[ŋ]V[ŋ] to the A[m]V[ŋ] stimulus *invariably* leads to a decrease in ‘ng’ responses, and an increase in two response types: both ‘m’, a response corresponding to the auditory component of the stimulus, and ‘n’, the fusion response. This is the case in all four conditions, so the English language participants $\times$ word-initial [ŋ] condition is no different from the other three conditions in this regard. Also of note is that, despite variations in the incidence of ‘n’ responses to V[ŋ] and A[ŋ]V[ŋ] stimuli, the absolute percentage of ‘n’ responses to the A[m]V[ŋ] stimulus is almost exactly the same in both the two initial conditions (28.13% for Thai, 27.5% for Australian English), and the two final conditions (58.33% for Thai, 56.25 for Australian English). This is consistent with, but by no means proving of course, the action of similar processes in Thai and English language participants in response to the A[m]V[ŋ] ‘McGurk’ stimulus.

3.6. Reaction Times on V[ŋ], A[ŋ]V[ŋ], and A[m]V[ŋ] Test Trials

Reaction times for all stimuli involving a visual [ŋ] are shown in Fig. 2. The results for V[ŋ], A[ŋ]V[ŋ] and A[m]V[ŋ] are discussed in turn.

3.6.1. V[ŋ] Trials

English subjects’ combined reaction times (RTs) on V[ŋ] trials were significantly longer than those of Thais, $F(1, 46) = 8.26$ ($M_{\text{difference}} = 201$ ms), but there were no other main or interaction effects. English speakers generally took longer to process visually ambiguous [ŋ], presumably due to general lack of experience with this viseme.

3.6.2. A[ŋ]V[ŋ] Trials

RTs for both language groups were generally slower on final than initial conditions for A[ŋ]V[ŋ] trials ($M_{\text{difference}} = 240$ ms), $F(1, 46) = 7.72$, presumably due to RTs being measured from sound onset resulting in slower RTs for final consonants. English were generally slower than Thai language participants ($M_{\text{difference}} = 224$ ms), $F(1, 46) = 21.71$, but there was no interaction with initial/final position, indicating that the slower English RTs are probably due to the general paucity of [ŋ] in the English language (Roberts, 1965), rather than the specific phonotactic illegality of initial [ŋ].

Figure 2. Reaction Times (RTs) for ‘n’ responses in ms to VO [ɲ] AV [ɲ] and A[m]V[ɲ] for in Initial and Final position for English and Thai participants. Means and standard errors.

3.6.3. *A[m]V[ɲ] Trials*

Despite the above differences on V[ɲ] and A[ɲ]V[ɲ] trials, there were *no significant differences between RTs on A[m]V[ɲ] trials* on any factor. Thus the McGurk combination of auditory [m] with visual [ɲ] requires equal processing time irrespective of position or native language. This further supports the notion that the A[m]V[ɲ] McGurk effect occurs in the same way across English and Thai, and for initial and final consonants despite the phonotactic illegality of [ɲ] in initial position in English.

3.7. *Discussion*

When [ɲ] is presented in the initial position, either in V[ɲ] or A[ɲ]V[ɲ], English language participants are affected by the visual ambiguity of V[ɲ] and show a native phonemic bias towards ‘n’ over ‘ng’ responses for initial [ɲ] and this is clear in both the incidence of ‘n’ responses and reaction times for ‘n’ responses. Such effects are not present either when these English language participants are presented with [ɲ] in the final position or when Thai language participants are presented with [ɲ] in either position.

Nevertheless, when the incongruent A[m]V[ɲ] is presented, in English language participants these language-specific and word-position-specific effects disappear; there is equal incidence of ‘n’ fusion responses by Thai and English language participants in the initial position and in the final position, and the reaction times are also equivalent for the two language groups. It is as if the illegality of [ɲ] in the initial position for English language participants has ceased to exist. These results are consistent with the integration of A[m] and

V[ŋ] in a space in which language-specific phonotactic constraints indeed do not exist, in a phonetic language-general space.

4. Experiment 2: Phonemic Repertoire Differences — Japanese vs. English

The equivalence of the McGurk effect across languages irrespective of phonotactic differences in Experiment 1 could be taken to show that auditory–visual integration occurs phonetically. However, it could be argued that as [ŋ] is a member of the English phoneme repertoire, albeit phonotactically illegal in initial position, English language speakers may have treated [ŋ] as a native phoneme even though it was in an illegal position, and the equivalent McGurk effects may then be said to have occurred in a phonemic space. The different pattern of responses to V[ŋ], A[ŋ]V[ŋ], and A[m]V[ŋ] make this unlikely. Nevertheless, the phonetic-not-phonemic conclusion would be strengthened if equivalent cross-language McGurk effects were found when a particular phone is a phoneme in one language but not in the other.

The next set of experiments rests on three established facts and phenomena: (1) the phone [ð], as in ‘*that*’ is phonemic in English but not Japanese, (2) the incongruent A[b]V[g] stimulus presented to English language speakers results predominantly in ‘d’ or ‘th’ responses, and (3) the relative incidence of different non-auditory responses when presented with incongruent AV stimuli differs as a function of vowel context, and (4) English language speakers’ relative incidence of ‘d’ and ‘th’ responses to A[b]V[g] stimulus differs as a function of whether the A[b]V[g] stimulus is A[ba]V[ga] or A[bi]V[gi]. Following a brief elaboration of these points, the results of two experiments with (i) English language and (ii) Japanese language participants are presented.

Hampson *et al.* (2003) investigated McGurk effect responses by English language participants across various consonant-vowel contexts. They found that with the [i] vowel, A[bi]V[gi] resulted in a majority of ‘di’ responses (87%) compared with 4% ‘gi’ responses; with the [a] vowel, A[ba]V[ga] resulted in equal numbers of ‘da’ and ‘ga’ responses (both 40%); and with the [u] vowel there was a complete reversal of the [i] vowel results — a majority of ‘gu’ responses (69%) compared with 13% ‘du’ responses.

Similar effects occur for Japanese participants with A[b]V[g] across vowel contexts. In addition, focussing on responses to A[b]V[g] across vowel contexts, Shigano found there were considerably more auditory-based ‘b’ responses than are found with English language participants (see also Sekiyama and Burnham, 2008), and that over and above this, across vowel contexts, [i] to [a] to [u], the incidence of auditory-based ‘b’ responses increased (38% to 43% to 67%, respectively) while the incidence of ‘d’ fusion responses decreased (59% to 30% to 3%, respectively).

Finally, Green (Green, 1996, 1998; Green and Norrix, 1997) has shown with English language participants that the relative distribution of ‘d’ and ‘th’ responses to the A[b]V[g] McGurk stimulus differs across [a] vs. [i] vowel contexts: English perceivers presented with A[ba]V[ga] typically provide more ‘tha’ than ‘da’ responses, whereas for A[bi]V[gi] there are typically more ‘di’ than ‘thi’ responses. It is this specific effect that is exploited in the two experiments to follow. As there is no [ð] in Japanese phonology, if the McGurk effect occurs at a phonetic level, then this phonetic vowel context, [a]/[i], by response, ‘d’/‘th’, effect should be apparent in both English and Japanese. That is there should be more ‘tha’ than ‘da’ responses for A[ba]V[ga] and more ‘di’ than ‘thi’ responses for A[bi]V[gi].

Note that this is different to the way in which Werker *et al.* (1992) exploited the lack of [ð] in French; there they investigated the incidence of ‘tha’ (compared with ‘ta’ or ‘da’) responses to A[ba]V[ð̥a] in English and French language participants. That design is still allows phonological influences to override any phonetic-level effects. In the design here it is the [a]/[i] × ‘d’/‘th’ *pattern of results* that is of interest and this should be influenced only by phonetic interactions, that is in A[b]V[g] the different co-articulatory effects of auditory [b] with [a] or [i] and visual [g] with [a] or [i].

This proposition was tested with a group of Australian English language participants, and three groups of Japanese participants with Beginner, Intermediate, or Advanced level of English proficiency. In Experiment 2a an English speaker presented the stimuli, and in Experiment 2b a Japanese speaker presented the stimuli.

4.1. Experiment 2a: the [a]–[i]/‘d’–‘th’ Effect With an English Speaker Method

Sixteen native adult Australian English speakers, and 48 native adult Japanese speakers were tested, 16 at each of three levels of English proficiency, Beginner, Intermediate and Advanced. All Japanese participants had had standard English lessons for six years in secondary school.

Beginner and Intermediate participants were volunteers from language schools in Sydney. Their English ability was determined at the time of enrolment on the basis of an oral interview and written test (National ELICOS, English Language Intensive Courses for Overseas Students, Accreditation Scheme, used in all accredited language schools). In the Beginner group, there were 11 females and five males (mean age = 24.5 years, range = 17.5–28 years), who had been in Australia for a mean of six weeks (range 1 week–3.5 months) prior to testing. In the Intermediate group there were seven females and nine males (mean age = 23.4 years, range = 17.8–31.5 years), who spent a mean period of three months (range 2 weeks–12 months) in Australia.

Advanced group participants were required to have worked in Australia or New Zealand for at least one but not more than 10 years and were recruited by advertisement and word of mouth. There were 12 females and four males (mean age = 28.4 years, range = 20–40.9 years), and they had lived in Australia for a mean of three years (range 1.1–8 years). The majority were undergraduate and postgraduate students studying at various institutions in Sydney and some were Japanese lecturers/teachers working in Sydney.

In the Australian English group there 10 females and six males (mean age = 24.0 years, range = 19–48 years) recruited from the first-year psychology student pool at the University of NSW. All but two of the English participants were monolingual, however both bilingual subjects reported speaking mostly English on a daily basis.

Stimuli were prepared by videorecording the head and shoulders of a female native Australian English speaker. She recorded the syllables [ba:], [bi:], [ga:], [gi:], [da:], [di:], [ða:], [ði:], [bga:], [bgi:], [gba:], [gbi:]. For the consonant clusters, she was asked to insert the schwa vowel, /ə/, between consonants, viz, [bəga:], [bəgi:], [gəba:] and [gəbi:].

The same procedure as in Experiment 1 was used to create the following stimuli: AO, VO, and AV matching presentations of the syllables [bV], [gV], [dV], [ðV], [bgV] and [gbV] and mismatching presentations of A[bV]V[gV] and A[gV]V[bV] (where V = vowel, [a] or [i]). Each trial lasted 4 s, with 1 s of black background intervening between trials. For VO and AV trials this consisted of 1 s of a motionless face, about 1 s of articulation, and 2 s of neutral expression. For the AO trials, the speaker's motionless face was presented for 4 s overdubbed with a speech sound.

Each vowel condition was presented separately. In each there were 18 practice trials, one of each of the six syllables [bV], [gV], [dV], [ðV], [bgV] and [gbV], in each of the three modes, AO, VO, and AV. For each vowel condition, there were then two 32 trial test blocks. Each block consisted of exactly the same trial types with trial presentation order varied between blocks, and test block sequence counterbalanced between participants. In each block there were two AO, two VO and two AV matching presentations of each of [bV], [gV], [dV] and [ðV], and four each of mismatching A[bV]V[gV] (McGurk stimulus), and A[gV]V[bV] (combination stimulus).

The procedure for testing and data collection was exactly the same as for Experiment 1 except that the response pad with its central 'ready' key had six response buttons (labelled 'b', 'g', 'd', 'th', 'bg' and 'gb') arranged in a semicircle around it.

4.2. Results and Discussion

The native phoneme bias results are presented then those for the incongruent trials.

Figure 3. Percent 'th' responses on AV, AO and VO [ð] trials. Experiment 2a, Australian English Speaker. Means and standard errors.

4.2.1. Native Phoneme Bias

Figure 3 shows the percentage of correct ('th') responses on AO, VO, and AV trials, A[ð], V[ð], A[ð]V[ð], collapsed over vowel context. As expected (because [ð] is not phonemically relevant in Japanese) English language participants made more correct responses than Japanese perceivers, and there was an increase in correct responses as the Japanese perceivers' experience with English increased (Beginner to Intermediate to Advanced). Statistical analyses confirmed these observations. Japanese participants made more errors than Australian English subjects on [ð] trials in all three modes, AV, $F(1, 60) = 11.16$, AO, $F(1, 60) = 24.26$, and VO, $F(1, 60) = 7.39$ trials. Additionally, correct responses for Japanese participants improved linearly as a function of English language experience in AV, $F(1, 60) = 8.03$, and AO trials, $F(1, 60) = 10.04$, and quadratically in VO trials, $F(1, 60) = 7.43$.

Thus, as would be expected on the basis of their native phonology, Japanese participants have a native phoneme bias against reporting 'th' when presented with [ð] in auditory-only, visual-only, and auditory–visual conditions. And, as would also be expected, this bias is ameliorated as a function of increasing facility with English.

4.2.2. Response to Incongruent A[b]V[g] Trials

Given this native phoneme bias, the next question concerns the pattern of 'd' and 'th' responses across [a] and [i] vowel contexts for the Japanese and English language participants.

Figure 4 shows the percentage of 'd' and 'th' responses to the A[b]V[g] stimulus in each vowel context, [a] and [i] — A[ba]V[ga] and A[bi]V[gi] combined across language groups; Fig. 5 shows the same separately for each language group. ANOVA of the two types of fusion responses, 'd' and 'th' in

Figure 4. Effect of Vowel Context on the Incidence of ‘d’ and ‘th’ fusion responses for all participants in Experiment 2a. Australian English speaker, means and standard errors.

Figure 5. Effect of Vowel Context on the Incidence of ‘d’ and ‘th’ fusion responses in Experiment 2a for each language group.

each vowel context, [a] and [i] revealed that English participants (83.6%) gave more fusion responses than did Japanese participants (70.3%), $F(1, 60) = 5.11$ (as would be expected — Sekiyama and Tohkura, 1991). There were generally more fusion responses in the [i] than [a] vowel environment (total = 84.1%, and 69.8% respectively), $F(1, 60) = 31.31$, but this difference was due mainly to the Japanese participants (total = 83.6% vs. 56.8% for [i] and [a], respectively) rather than the English language participants (total = 84.4% vs. 82.8%), $F(1, 60) = 8.9$. In addition, as expected on the basis of a native phoneme bias, Japanese made many more ‘da’ (42.3%) than ‘tha’ (28.0%) responses whereas English language participants made more ‘tha’ (73.0%) than ‘da’ (10.5%) responses, $F(1, 60) = 36.25$.

Table 3.
Acceptance/rejection of hypotheses for ‘d’ responses in [a] and [i] vowel contexts and ‘th’ responses in [a] and [i] vowel contexts for each of the four language groups in Experiment 2a (English language stimulus person) and Experiment 2b (Japanese language stimulus person)

Experiment and speaker	Response and hypothesis	Japanese participants			Australian English participants
		Beginners	Intermediate	Advanced	
2a English speaker	‘d’	✓	✓	✓	✓
	[i] > [a]				
	‘th’	✓	×	✓	✓
	[a] > [i]				
2b Japanese speaker	‘d’	✓	✓	✓	✓
	[i] > [a]				
	‘th’	✓	✓	0	×
	[a] > [i]				

Of particular interest is the [a]/[i] vowel × ‘d’/‘tha’ response effect. There was a significant overall [a]/[i] vowel × ‘d’/‘tha’ response effect over all four groups, $F(1, 60) = 10.80$, and despite the above differences between Japanese and English language participants in the number of fusion responses and the relative proportion of ‘d’ and ‘th’ responses, this overall [a]/[i] vowel × ‘d’/‘th’ response effect did not interact with language background, $F(1, 60) = 0.058$. As can be seen in Fig. 4 the overall effect is due to more ‘th’ than ‘d’ responses in [a] vowel context, and more ‘d’ than ‘th’ responses in [i] vowel context, exactly as was hypothesised. For McGurk stimuli, English language and Japanese language participants respond to changes in vowel context in a statistically equivalent manner, presumably at a phonetic level of processing.

Although there was no interaction with language background, it is of interest to inspect the individual [a]/[i] vowel × ‘d’/‘th’ plots (see Fig. 5). The effect can be viewed in two ways (and these are set out in graphic form in Table 3). First, in terms of the ‘d’ responses in [a] and [i] vowel contexts, all four groups showed the expected effect — more ‘d’ responses in [i] than [a] vowel contexts. Second, in terms of the ‘th’ responses in [a] and [i] vowel contexts, three of the four groups showed the expected effect — more ‘th’ responses in [a] than [i] vowel contexts. The Intermediate Japanese group showed a reversal of the effect and for the Advanced Japanese group the advantage was quite small. These departures are not significant, as there was no interaction of the [a]/[i] vowel × ‘d’/‘th’ with language, but nevertheless, it could be argued that a more powerful experiment with more participants may be required to detect any language background differences. However there was sufficient power in

the analyses to detect significant language background differences in other respects: for Japanese vs. English language participants there were less fusion responses overall; more fusion responses in the [i] than [a] vowel environment; and more ‘da’ than ‘tha’ responses overall. Nevertheless, further elaborations would be useful.

4.3. *Experiment 2b: the [a]–[i]–‘d’–‘th’ Effect’ With a Japanese Speaker*

The results in Experiment 2a, and the veracity of the conclusion that the same or similar phonetic effects occur for Japanese and English language participants, may rest on the fact that an English speaker was used and that both English and Japanese language participants would expect an English language speaker to use the [ð] consonant. A Japanese speaker would not be expected to do so, and this may change the nature of any fusion responses. Experiment 2b was the same as Experiment 2a except that a Japanese speaker was used as the stimulus person.

Four new groups of 16 participants were tested. In the Beginner group there were 12 females and four males (mean age = 24.1 years, range = 18–31 years) and they had been in Australia for a mean of 4.2 weeks (range: 3–11 weeks); the Intermediate group consisted of 12 females and four males (mean age = 24.1 years, range = 18–31 years) and they had been in Australia for a mean of 4.1 months (range: 2–8 months); and the Advanced group had 11 females and five males (mean age = 26.6 years, range = 20–37 years), and they had been in Australia for a mean of 4.13 years (range: 1.5–9.3 years). The Australian English group, 11 females and five male participants (mean age = 20.7 years, range = 18–26 years) were recruited from the first-year psychology student pool at the University of NSW.

4.4. *Results and Discussion*

4.4.1. *Native Phoneme Bias*

Figure 6 shows the percentage of ‘th’ responses on AV, AO, and VO [ð] trials. As can be seen the results on [ð] trials with a Japanese speaker were similar to when an Australian English speaker was used, but there were some differences. Both English (mean = 95.31%) and Japanese (mean = 88.28%) participants performed well on AV [ð] trials, and there was no significant difference between them. Similarly when only visual information for [ð] was available, both English (mean = 94.54%) and Japanese (mean = 85.15%) participants performed well, but the ability to respond to VO [ð] correctly with ‘th’ increased for Japanese participants as a function of their experience with English, $F(1, 60) = 4.22$. With only auditory information, however, there was a clear advantage for Australian English over Japanese perceivers in their perception of ‘th’ from AO [d] (mean = 68.75% vs. 49.05%), $F(1, 60) = 16.62$,

Figure 6. Percent ‘th’ responses on AV, AO and VO [ð] trials, Experiment 2b, Japanese speaker. Means and standard errors.

Figure 7. Effect of Vowel Context on the Incidence of ‘d’ and ‘th’ fusion responses for all participants, Japanese Speaker, Experiment 2b. Means and standard errors.

indicating a native phoneme bias in the Japanese listeners when only auditory information is available.

4.4.2. Response to Incongruent A[b]V[g] Trials

There was a similar native phoneme bias here for Japanese participants perceiving a Japanese speaker as in the previous experiment when an Australian English speaker was used. Was there the same phonetic effect of the [a] vs. [i] vowel context? The results for the A[b]V[g] ‘McGurk’ trials collapsed over participant groups are shown in Fig. 7; Fig. 8 shows the same separately for each language group.

Compared to Experiment 2a with an English language speaker, here with a Japanese speaker, the percentage of ‘th’ responses was greatly reduced such that there were far fewer ‘d’ than ‘th’ responses for both the Japanese *and* the

Figure 8. Effect of Vowel Context on the Incidence of ‘d’ and ‘th’ fusion responses in each language group. Japanese speaker, Experiment 2b. Means and standard errors.

English language participants, $F(1, 60) = 335.95$. This was accompanied by a departure from the usual ‘Japanese McGurk effect’ (Sekiya and Burnham, 2008; Sekiya and Tohkura, 1991) — the number of fusion responses was statistically equivalent between Japanese (mean = 56.5%) and English language participants (mean = 62.5%), $F(1, 60) = 1.33$. Additionally, as in Experiment 2a, there were generally more fusion responses in the [i] than [a] vowel environment, $F(1, 60) = 149.81$, and while this was a greater difference for the Japanese participants ($M_{\text{difference}} = 86.98\%$ vs. 26.04% for [i] and [a], respectively), $F(1, 60) = 9.72$, the difference was also large here for the English language participants (mean = 77.34% vs. 47.66% for [i] and [a], respectively).

Of particular interest is the [a]/[i] vowel $\times$ ‘d’/‘th’ response effect. The overall [a]/[i] vowel $\times$ ‘d’/‘th’ response effect was significant, $F(1, 60) = 121.37$. There were as hypothesised, more ‘d’ than ‘th’ responses in the [i] vowel context, but there were not more ‘th’ than ‘d’ responses in the [a] vowel context. This is presumably due to the relative dearth of ‘th’ responses when a Japanese speaker presented the stimuli.

Here this [a]/[i] vowel $\times$ ‘d’/‘th’ response effect, unlike in Experiment 2a, interacted with language background, $F(1, 60) = 14.00$. As can be seen in Fig. 8 (and in the graphic representation in Table 3), there were, as expected more ‘d’ responses in the [i] than the [a] vowel context. So half of the expected results were obtained. With respect to the ‘th’ responses there was minimal support for more in the [a] than [i] vowel context for Beginner and Intermediate Japanese learners of English, no ‘th’ responses at all for the

Japanese Advanced, and a reversal of the expected effect for the English language group. As even the English language group did not show the expected result, and as there were very few ‘th’ responses for the Japanese groups, no meaningful statements can be made about the ‘th’ response half of the expected results. Nevertheless, all four groups showed the same expected effect for the ‘d’ responses

Thus there is evidence for similar phonetic level processing by the English and the Japanese participants, but here the evidence for the full [a]/[i] vowel × ‘d’/‘tha’ response effect is not as strong as in Experiment 2a as it appears that the Japanese speaker did not afford the perception of a ‘th’ consonant from A[b]V[g] even for English language participants. It is perhaps the case that the particular acoustics of Japanese [b], or the particular visible articulatory movements of Japanese speech, or both affect perception of ‘th’. Resolution of this issue awaits further research.

5. General Discussion

In the introduction behavioural and brain response data were presented to show where, how, and when and in what linguistic and non-linguistic auditory–visual integration occurs, both in artificial (McGurk) contexts or more natural contexts. However, it was concluded that whether the interaction between auditory and visual percepts of speech occur at a language-general, phonetic, or the language-specific, phonemic level, is yet to be determined. While further research employing brain response and behavioural identification data would clarify integration of multisensory speech stimuli, the research studies reported here provide new information, and may assist in guiding such further research.

Experiments 1 and 2 provide evidence that in the McGurk effect, auditory and visual speech information is initially integrated at a phonetic level of processing. Experiment 1 provides such evidence despite a difference in phonotactic constraints in Thai and English, and Experiment 2 provides such evidence despite a difference in phoneme response repertoire in English and Japanese.

In the Thai–English study (Experiment 1), despite Australian English participants’ tendency to perceive initial visual [ŋ] and auditory–visual [ŋ] as ‘n’ more than do Thai participants, in response to the A[m]V[ŋ] McGurk stimulus this phonotactically-based bias is no longer evident. Australian English participants perceive an equivalent number of ‘n’ fusions to A[m]V[ŋ] as their Thai counterparts in both word-initial and word-final conditions.

In the Japanese–English study with an English language speaker (Experiment 2a) there is a general crossover effect in the incidence of ‘d’ and ‘th’ in [a] and [i] contexts: there are relatively more ‘tha’ than ‘da’ responses to

A[ba]V[ga], and relatively more ‘di’ than ‘thi’ responses to A[bi]V[gi]. This shows a phonetic level of integration. There was no significant interaction of this effect with language group, and inspection of each group reveals similar results across groups with minor deviations. When this experiment was repeated with a Japanese speaker (Experiment 2b), there was a relative absence of ‘th’ responses, even for the English language participants; and the hypothesis was only supported with respect to the pattern of ‘d’ responses across the two vowel contexts.

The Japanese results also reveal other influences on auditory–visual speech perception, and may inform what has been called the Japanese McGurk effect — fewer fusion responses by Japanese perceivers. The number of fusion responses is greater in the [i] vowel context, and this is particularly so for Japanese participants looking at a Japanese speaker. This suggests that there is a subtle interplay of phonetic, phonemic, and speaker variables in the McGurk effect. Phonetically, with the [a] vowel the phonetic conditions are more conducive to the perception of ‘th’. Phonemically, ‘th’ is irrelevant for Japanese perceivers. And with regard to the speaker, Japanese speakers’ articulation may support the perception of ‘d’ rather than ‘th’, even for English perceivers. Thus it is possible that the ‘Japanese McGurk effect’ (Sekiya and Burnham, 2008; Sekiya and Tohkura, 1991) is in part the product of the common use of the [a] vowel in McGurk effect studies (also see Shigeno, 2000).

The results from these studies show that low-level early, and with respect to speech, phonetic processes determine *what is perceived* when incongruous auditory and visual information is presented. Over and above this, there appears to be later, phonemic and even cultural effects on *what is reported* when incongruous auditory and visual information is presented. We contend that auditory–visual integration occurs in a common representational space with a motoric (Robert-Ribes *et al.*, 1995, 1996), articulatory/gestural (Lieberman and Mattingly, 1985, 1989; Studdert-Kennedy and Goodell, 1992), or phonetic (Dodd and Burnham, 1988; Green and Kuhl, 1991) basis. Whatever the nature of this common metric, the important point is that auditory–visual integration occurs early and directly, devoid of any influence of learned associations or phonological prototypes. Certain visual and auditory information may be clearer with some speakers than others, so at this early stage there will be varying degrees of information available for integration. Following integration into phonetic categories, there can be late post-categorical effects of native phoneme inventory or even phonotactics, and cross-cultural factors. Thus a distinction can be made between the *direct* McGurk effect (phonetic-level auditory–visual integration), and the *reported* McGurk effect, additionally influenced by later phonological (post-categorical) effects.

Much has been learned about the world from studying the conditions under which a particular system breaks down in disciplines as diverse as engineering,

medical science and psychology. The McGurk effect entails such a breakdown — in this case in auditory–visual speech perception. In most perception studies, e.g., in illusions such as the Poggendorf illusion (Day and Dickinson, 1976), the breakdown is thought to occur due to going beyond the normal limits of a system or misapplying a usually useful perceptual strategy. However, it has been suggested that the (neural) processes involved in the McGurk effect and normal integration of auditory and visual speech information may be different (see Introduction, and Alsius *et al.* in this issue). If so then it is possible that the discovery of auditory–visual fusions by McGurk and MacDonald in 1976 (see MacDonald in this issue) have led to erroneous conclusions regard auditory–visual speech perception.

Such an extreme conclusion is not warranted, for at the very least the McGurk effect (along with other parallel movements in the *Zeitgeist*) has led to increased research attention to auditory–visual and more generally inter-modal speech and other perception. In addition, the McGurk effect may just tell us something new. Consider as an example the results of Experiment 1 here. Following the above ideas of early direct fusion and later post-categorical effects, with congruent auditory–visual presentations of A[ɪ]V[ɪ] in the initial position there is integration, and then phonemic/phonotactic processes come into play such that English language perceivers report ‘n’ more often, and do so with slower processing time than do Thai language perceivers. However, when the incongruent A[m] replaces A[ɪ], it appears that the immediate engagement of phonemic post-categorical processes is blocked; perception of A[m]V[ɪ] appears to employ similar processes (equal reaction times) and results in similar responses (equivalent incidence of ‘n’ for English and for Thai language perceivers). Such a phenomenon may assist in the understanding of phonetic and phonemic processing — when there is conflict or ambiguity, the processes learned as a product of specific linguistic experience (phonemic processing) can be bypassed in order to access more basic levels of perception (see Burnham, Tyler and Horlyck, 2002). In the McGurk effect this bypass results in an illusion — but maybe such a bypass may lead to more veridical perception in other circumstances, such as in second language learning when naturally-occurring pairings of familiar mouth movements with unfamiliar speech sounds or unfamiliar mouth movements with familiar speech sounds.

Acknowledgements

The assistance and expertise of Shelia Keane and Megan Smith in preparing and running experiments, John Fowler and Michelle Nicol in programming, and Amanda Reid in some early editing is gratefully appreciated. Parts of the

studies reported here were previously reported in a chapter (Burnham, 1998) and a conference contribution (Burnham and Keane, 1997).

References

- Alsius, A., Paré, M. and Munhall, K. G. (2017). Forty years after Hearing lips and seeing voices: the McGurk effect revisited, *Multisens. Res.* DOI:10.1163/22134808-00002565.
- Baart, M., Stekelenburg, J. J. and Vroomen, J. (2014). Electrophysiological evidence for speech specific audiovisual integration, *Neuropsychologia* **53**, 115–121.
- Bernstein, L. E., Burnham, D. and Schwartz, J.-L. (2002). Special session: Issues in audiovisual spoken language processing (When, Where, and How?), in: *7th International Conference on Spoken Language Processing*, Denver, CO, USA. pp. 1445–1448. ISBN: 1876346-40-X.
- Burnham, D. (1998). Language specificity in the development of auditory–visual speech perception, in: *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory–Visual Speech*, R. Campbell, B. Dodd, D. Burnham (Eds), pp. 27–60. Psychology Press, Hove, UK.
- Burnham, D. and Dodd, B. (1996). Auditory–visual speech perception as a direct process: the McGurk effect in human infants and across languages, in: *Speechreading by Humans and Machines*, D. G. Stork and M. E. Hennecke (Eds), pp. 103–114. Springer-Verlag, Berlin, Germany.
- Burnham, D. and Dodd, B. (1998). Familiarity and novelty in infant cross-language studies: factors, problems, and a possible solution, in: *Advances in Infancy Research*, Vol. 12, C. Rovee-Collier, L. Lipsitt and H. Hayne (Eds), pp. 170–187. Ablex Publishing Corporation, Stanford, CT, USA.
- Burnham, D. and Dodd, B. (2004). Auditory–visual speech integration by pre-linguistic infants: perception of an emergent consonant in the McGurk effect, *Dev. Psychobiol.* **44**, 204–220.
- Burnham, D. and Keane, S. (1997). The Japanese McGurk effect: the role of linguistic and cultural factors in auditory–visual speech perception, in: *Proceedings of the Workshop on Auditory–Visual Speech Processing: Cognitive and Computational Approaches*, Rhodes, Greece, pp. 93–96.
- Burnham, D., Tyler, M. and Horlyck, S. (2002). Periods of speech perception development and their vestiges in adulthood, in: *An Integrated View of Language Development: Papers in Honor of Henning Wode*, P. Burmeister, T. Piske and A. Rohde (Eds), pp. 281–300. Wissenschaftlicher Verlag Trier, Trier, Germany.
- Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., Woodruff, P. W. R., Iversen, S. D. and David, A. S. (1997). Activation of auditory cortex during silent lipreading, *Science* **276**, 593–596.
- Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G., McGuire, P., Suckling, J., Brammer, M. J. and David, A. S. (2001). Cortical substrates for the perception of face actions: an fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning), *Cogn. Brain Res.* **12**, 233–243.
- Choi, J., Cutler, A. and Broersma, M. (2017). Early development of abstract language knowledge: evidence from perception–production transfer of birth-language memory, *R. Soc. Open Sci.* **4**, 160660. DOI:10.1098/rsos.160660.

- Day, R. H. and Dickenson, R. G. (1976). The components of the Poggendorff illusion, *Br. J. Psychol.* **67**, 537–552.
- Desjardins, R. N. and Werker, J. F. (2004). Is the integration of heard and seen speech mandatory for infants? *Dev. Psychobiol.* **45**, 187–203.
- Dodd, B. and Burnham, D. (1988). Processing speechread information, *Volta Rev. N. Reflect. Speechread.* **90**, 45–60.
- Green, K. (1996). The use of auditory and visual information in phonetic perception, in: *Speechreading by Humans and Machines*, D. G. Stork and M. E. Hennecke (Eds), pp. 55–77. Springer-Verlag, Berlin, Germany.
- Green, K. (1998). The use of auditory and visual information during phonetic processing: implications for theories of speech perception, in: *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory–Visual Speech*, R. Campbell, B. Dodd and D. Burnham (Eds), pp. 3–25. Psychology Press, Hove, UK.
- Green, K. P. and Kuhl, P. K. (1989). The role of visual information in the processing of place and manner features in speech perception, *Percept. Psychophys.* **45**, 34–42.
- Green, K. P. and Kuhl, P. K. (1991). Integral processing of visual place and auditory voicing information during phonetic perception, *J. Exp. Psychol. Hum. Percept. Perform.* **17**, 278–288.
- Green, K. P. and Miller, J. L. (1985). On the role of visual rate information in phonetic perception, *Percept. Psychophys.* **38**, 269–276.
- Green, K. P. and Norrix, L. W. (1997). Acoustic clues to place of articulation and the McGurk effect: the role of release bursts, aspiration, and formant transitions, *J. Speech Lang. Hear. Res.* **40**, 646–665.
- Green, K. P., Kuhl, P. K., Meltzoff, A. N. and Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: female faces and male voices in the McGurk effect, *Percept. Psychophys.* **50**, 524–536.
- Hampson, M., Guenther, F., Cohen, M. and Nieto-Castanon, A. (2003). Changes in the McGurk effect across phonetic contexts, Technical Report CAS/CNS 03-006, Boston University, MA, USA. Retrieved from <https://pdfs.semanticscholar.org/dc98/5ea3a57d1d1175b0a4ab595e6649de409e9a.pdf>.
- Kuhl, P. K. and Meltzoff, A. N. (1984). The intermodal representation of speech in infants, *Infant Behav. Dev.* **7**, 361–381.
- Kuhl, P. K. and Meltzoff, A. N. (1988). Speech as an intermodal object of perception, in: *Perceptual Development in Infancy: the Minnesota Symposia on Child Psychology*, Vol. 20, A. Yonas (Ed.), pp. 235–266. Erlbaum, Hillsdale, NJ, USA.
- Liberman, A. M. and Mattingly, I. G. (1985). The motor theory of speech perception revised, *Cognition* **21**, 1–36.
- Liberman, A. M. and Mattingly, I. G. (1989). A specialization for speech perception, *Science* **243**, 489–494.
- MacDonald, J. (2017). Hearing lips and seeing voices: the origins and development of the ‘McGurk Effect’ and reflections on audio–visual speech perception over the last 40 years, *Multisens. Res.* DOI:10.1163/22134808-00002548.
- MacDonald, J. and McGurk, H. (1978). Visual influences on speech perception processing, *Percept. Psychophys.* **24**, 253–257.

- Massaro, D. (1987). *Speech Perception by Eye and Ear: the Psychology of Lip-Reading*. Erlbaum, London, UK.
- Massaro, D., Thompson, L., Barron, B. and Laren, E. (1984). Developmental changes in visual and auditory contributions to speech perception, *J. Exp. Child Psychol.* **41**, 93–113.
- Massaro, D. W. (1996). Bimodal speech perception: a progress report, in: *Speechreading by Humans and Machines*, D. G. Stork and M. E. Hennecke (Eds), pp. 79–101. Springer-Verlag, Berlin, Germany.
- McGurk, H. and Buchanan, L. (1981). Bimodal speech perception: vision and hearing. Unpublished manuscript, Department of Psychology, University of Surrey, Guildford, UK.
- McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices, *Nature* **264**, 746–748.
- Mills, A. E. (1987). The development of phonology in the blind child, in: *Hearing by Eye: the Psychology of Lip-Reading*, B. Dodd and R. Campbell (Eds), pp. 145–161. Erlbaum, Hove, UK.
- Mills, A. E. and Thiem, R. (1980). Auditory–visual fusions and illusions in speech perception, *Linguist. Ber.* **68/80**, 85–108.
- Munhall, K. G., Gribble, P., Sacco, L. and Ward, M. (1996). Temporal constraints on the McGurk effect, *Percept. Psychophys.* **58**, 351–362.
- Musacchia, G., Sams, M., Nicol, T. and Kraus, N. (2006). Seeing speech affects acoustic information processing in the human brainstem, *Exp. Brain Res.* **168**, 1–10.
- Robert-Ribes, J., Schwartz, J.-L. and Escudier, P. (1995). Auditory, visual, and audiovisual vowel representations: experiments and modelling, in: *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Vol. 3, K. Elenius and P. Branderud (Eds), pp. 114–121. ICPHS and Stockholm University, Stockholm, Sweden.
- Robert-Ribes, J., Piquemal, M., Schwartz, J.-L. and Escudier, P. (1996). Exploiting sensor fusion architectures and stimuli complementarity in AV speech recognition, in: *Speechreading by Humans and Machines*, D. G. Stork and M. E. Hennecke (Eds), pp. 193–210. Springer-Verlag, Berlin, Germany.
- Rosenblum, L. D., Schmuckler, M. A. and Johnson, J. A. (1997). The McGurk effect in infants, *Percept. Psychophys.* **59**, 347–357.
- Saito, D. N., Yoshimura, K., Kochiyama, T., Okada, T., Honda, M. and Sadato, N. (2005). Cross-modal binding and activated attentional networks during audio–visual speech integration: a functional MRI study, *Cereb. Cortex* **15**, 1750–1760.
- Saldaña, H. M. and Rosenblum, L. D. (1993). Visual influences on auditory pluck and bow judgments, *Percept. Psychophys.* **54**, 406–416.
- Sekiyama, K. (1994). Differences in auditory–visual speech perception between Japanese and Americans: McGurk effect as a function of incompatibility, *J. Acoust. Soc. Jpn* **15**, 143–158.
- Sekiyama, K. and Burnham, D. (2008). Impact of language on development of auditoryvisual speech perception, *Dev. Sci.* **11**, 303–317.
- Sekiyama, K. and Tohkura, Y. (1991). McGurk effect in non-English listeners: few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility, *J. Acoust. Soc. Am.* **90**, 1797–1805.
- Sekiyama, K. and Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception, *J. Phon.* **21**, 427–444.
- Sekiyama, K., Braidia, L., Nishino, K., Hayashi, M. and Tuyo, M. (1995). The McGurk effect in Japanese and American perceivers, in: *Proceedings of the XIIIth International Congress*

- of *Phonetic Sciences*, Vol. 3, K. Elenius and P. Branderud (Eds), pp. 214–217. ICPhS and Stockholm University, Stockholm, Sweden.
- Shigeno, S. (2000). Influence of vowel context on the audio–visual speech perception of voiced stop consonants, *Jpn Psychol. Res.* **42**, 155–167.
- Studdert-Kennedy, M. (1986). Development of the speech perceptuomotor system, in: *Precursors of Early Speech*, B. Lindblom and R. Zetterstrom (Eds), pp. 205–217. Stockton Press, New York, NY, USA.
- Studdert-Kennedy, M. and Goodell, E. W. (1992). Gestures, features and segments in early child speech, *Haskins Laboratories Status Report on Speech Perception*, SR-111/112, 89–102.
- Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio–visual speech perception, in: *Hearing by Eye: the Psychology of Lip-Reading*, B. Dodd and R. Campbell (Eds), pp. 3–52. Erlbaum, Hillsdale, NJ, USA.
- Van Wassenhove, V., Grant, K. W. and Poeppel, D. (2007). Temporal window of integration in auditory–visual speech perception, *Neuropsychologia* **45**, 598–607.
- Van Wassenhove, V., Grant, K. W. and Poeppel, D. (2015). Visual speech speeds up the neural processing of auditory speech, *Proc. Natl Acad. Sci. USA* **102**, 1181–1186.
- Werker, J. F., Frost, P. and McGurk, H. (1992). Cross-language influences on bimodal speech perception, *Can. J. Psychol.* **46**, 551–568.