Auditory-Visual Perception of Thai Consonants by Thai and Australian Listeners

Denis Burnham
School of Psychology
University of New South Wales
Sydney 2033, Australia

Introduction Adults report perceiving /da/ when the auditory speech syllable /ba/ is dubbed onto the lip movements for /ga/ (McGurk & McDonald, 1976). A similar effect occurs when auditory /ma/ is dubbed onto /ŋa/ lip movements - /na/ is perceived. The occurrence of this McGurk effect or "fusion illusion" suggests that visual information is an integral part of speech perception, not just an adjunct to be used when auditory information is degraded or absent.

If this is the case then it is of interest to know whether linguistic experience influences the processing of auditory-visual speech or whether it is processed at a pre-phonological level (Dodd & Burnham, 1988). In this regard, two specific questions are of interest: firstly, at what age of development is integrated auditory-visual speech perception apparent, and secondly at what stage of information processing is auditory-visual speech perceived. For the first question, if young infants were shown to perceive the McGurk effect, then this would suggest that processing of auditory-visual speech information occurs at a pre-phonological level. For the second question, if the McGurk effect was found to occur for adults' with different language backgrounds, even when the auditory or visual component in the effect is phonologically irrelevant, then it could be said that adults process auditory-visual speech at a pre-phonological level.

There is some evidence bearing on the developmental issue. Recently, we tested 4-month-old infants for their ability to perceive the auditory /ba/ plus visual /ga/ effect. The experiment was constructed such that infants could see a woman speaking whenever they visually fixated the woman's face. Thus speech to the baby always involved auditory-visual input: infants never saw the lips move without a voice and they were not presented with the voice unless they fixated the face. Visual fixation

times to the face were recorded throughout the experiment. An experimental group was familiarized with a combination stimulus, auditory /ba/ plus visual /ga/, while a control group was familiarized with a matching stimulus, auditory /ba/ plus visual /ba/. When fixations of these stimuli habituated to a predetermined criterion, infants in both groups were given three test trials. In each test trial they could hear one stimulus, /ba/, /da/ or /ða/, whenever they fixated the woman's face. (The /ð/ phoneme was included because it is sometimes perceived as a result of auditory /b/ and visual /g/.) Results of visual fixations showed that infants in the /ba/ plus /ga/ experimental group perceived the /da/ and /ða/ stimuli to be the more familiar of the three test sounds, while infants in the /ba/ plus /ba/ control group perceived the /ba/ sound to be the most familiar test sound. This evidence for perception of the McGurk effect in infancy implies that processing of auditory and visual speech information occurs at a prephonological level because, while it is known that young infants discriminate various speech contrasts (Burnham et al., 1991), this is thought to occur at an accustic or phonetic level rather than at a phonological level (Burnham, 1986; Kuhl, 1978).

The results of this infant study show that auditory-visual perception of speech is possible in the absence of specific phonological experience, and provide important implications for the level at which auditory and visual speech information is combined. However, it is not clear what role phonological experience plays in auditory-visual speech perception when it is available. In order to address this issue, studies must be conducted in which subjects' phonological experience is systematically varied. This can be done by testing adults from different language backgrounds, for while young infants may perceive the universal set of phonetic contrasts, adults' perception of speech contrasts is constrained to some extent by the phonological distinctions made in their native language (Burnham 1986; Werker & Tees, 1984). In the current experiment, the fact that /ŋ/ is used in both initial and final positions in Thai but only in the initial position in English, will be exploited in order to investigate the effect of phonology on auditory-visual speech perception.

One previous study of cross-language auditory-visual speech perception has been conducted. Werker

et al. (1987) paired auditory /ba/ with visual /ba/, /da/, /va/ or /ð̆a/ and asked native English and native French speakers to identify what they perceived. The first three of these consonants are phonologically relevant in both languages while /ð̆/ is only relevant in English. It was found that the incidence of /ð̆a/ identifications for auditory /ba/ plus visual /ð̆a/ was greatest for the English speakers; and for the French speakers, /ð̆a/ identifications increased as a function of their English language experience. These results suggest some influence of phonology on auditory-visual speech perception. However, the reason that French subjects did not make as many /ð̆a/ responses as English subjects is not clear from the results of this study. On the one hand, French speakers may not have perceived /ð̆/ as the product of auditory /b/ and visual /ð̆/ because they had a phonologically determined bias against perceiving /ð̆/. On the other hand, they may have had a phonologically determined response bias against reporting /ð̆/, i.e., despite their perception of /ð̆/, French speakers may not have labelled their perceptual experience as /ð̆/ due to the absence of this phoneme in their native language. The current study is designed to overcome this difficulty.

In the current study Thai and English speakers were presented with auditory /m/ plus visual /ŋ/ auditory-visual combinations as well as matching, auditory-only, and visual-only conditions. These were presented in two ways: as final position consonants and as initial position consonants. To determine the role of phonological relevance in the McGurk effect, English speakers' perception of auditory /m/ plus visual /ŋ/ in the initial position is the condition of interest. Due to the phonological irrelevance of initial /ŋ/ in English and the ambiguity of the visual component (viseme) of /ŋ/ (Mills, 1987; Mills & Thiem, 1980), English subjects would be expected to respond "n" more often in the initial position than in the final position for visual-only /ŋ/ and matching auditory-visual /ŋ/. However, in the auditory /m/ plus visual /ŋ/ condition the outcome for "n" responses would depend on the level at which auditory-visual speech perception occurs. If it occurs at a pre-phonological level then English subjects' response pattern to auditory /m/ plus visual /ŋ/ trials for initial and final positions should not differ from that of the Thais. That is, both English and Thai speakers should be influenced by the phonetic possibilities (/n/, /ŋ/, /d/, /g/ etc.) which are said

to be specified by the lip movements for /ŋ/ (Mills, 1987; Mills & Thiem, 1980). If, however, auditory-visual processing occurs at a phonological level, then the relative incidence of "n" responses to auditory /m/ plus visual /ŋ/ should be greater in initial than final position for English but not Thai speakers.

Method

Subjects 48 adult subjects were tested, 24 native Thai speakers and 24 native English speakers. English speakers were all basically monolingual with no experience of Thai, or any language in which /ŋ/ is used initially. Thai speakers could also speak English reasonably well, however, this was not considered a problem as the principal manipulation of interest in the experiment was the phonological irrelevance of initial /ŋ/ for English speakers.

Apparatus and Stimulus Materials Subjects were tested at the University of NSW in a room containing a video monitor and a response key. The response key had a central button which served as a "ready" key and 5 response buttons in a semicircle, each 6 cm from the ready key. These were labelled 'm', 'm-ng', 'n', 'ng-m', and 'ng' for the English speakers, and 'ม','ม-ง','น', 'ง-ม , and ง for the Thai speakers. The experimenter sat in an adjacent room which also housed an IBM-compatable 386 computer used to control the equipment, present stimuli, and measure subjects' responses.

Auditory-visual stimuli were recorded from a female native Thai speaker. These were edited to produce the required stimuli: auditory only, visual only, and matching and mismatching auditory-visual stimuli. The visual component of the stimuli consisted of the original videorecordings presented on videotape. Auditory stimuli were digitized versions of the videotaped sounds and were presented directly from disk via filters and amplifier. Exact dubbing of sounds was achieved by using the original sound on one channel of the videotape as an input to the computer via a voice-key. This input triggered the computer to play a prearranged sound from disk. For example, for an auditory /m/ plus visual /ŋ/ trial the visual /ŋ/ was seen by the listener on the videomonitor. However, the original auditory component of /ŋ/ on the videotape was not audible to the listener - instead it was directed to the computer via a voice-key, the impulse from the voice key serving to trigger the computer to play auditory /m/ from disk through the

speaker on the videomonitor. This same impulse triggered a computer clock so that subjects' button press reactions were timed from the onset of the sound. Similar dubbing was conducted even in matching conditions to ensure conformity across conditions. In auditory alone conditions the speaker on the videotape was presented visually with no accompanying lip movements and the sound from disk and the computer clock were triggered by a tone on the second channel of the videotape. In visual alone conditions the original auditory component on the videotape was not audible to the subject but triggered the computer to play "silence" and to start the computer clock.

Procedure Subjects sat at a desk on which the monitor was situated approximately 50 cm from their eyes with the response key on the desk in front of the monitor. Each subject was tested individually. The experiment was divided into two phases, one in which consonants were presented in word-initial position and one in which they were presented in word-final position with order of presentation counterbalanced between subjects. In both phases the consonants were accompanied by an /a:/ vowel. In each phase there was a practice block and a test block.

The practice block consisted of 15 trials in which auditory and visual information always matched. Each of the five consonants or consonant clusters on the response key was presented three times, once in auditory-visual mode, once in auditory only mode, and once in visual only mode. These were /ma:/, /məŋa:/, /na:/, /ŋəma:/, and /ŋa:/ in the initial consonant phase , and /a:m/, /a:məŋ/, /a:n/, /a:ŋəm/, and /a:ŋ/ in the final consonant phase. These practice trials were included in order to (a) alert subjects to the possible responses they could make, (b) give them practice at using the five keys and (c) eliminate the data of subjects whose responses were too slow or inaccurate. Two criteria were employed for the exclusion of data: subjects were required (i) to respond within 2.5 secs of sound onset on at least 20 of the 30 practice trials (across initial and final phases); and (ii) respond correctly on at least 50% of valid trials in each phase, and across both phases. On the basis of these criteria the data of four Thai and five English speakers could not be used. Nevertheless, there were still 24 subjects in the final sample for each language group.

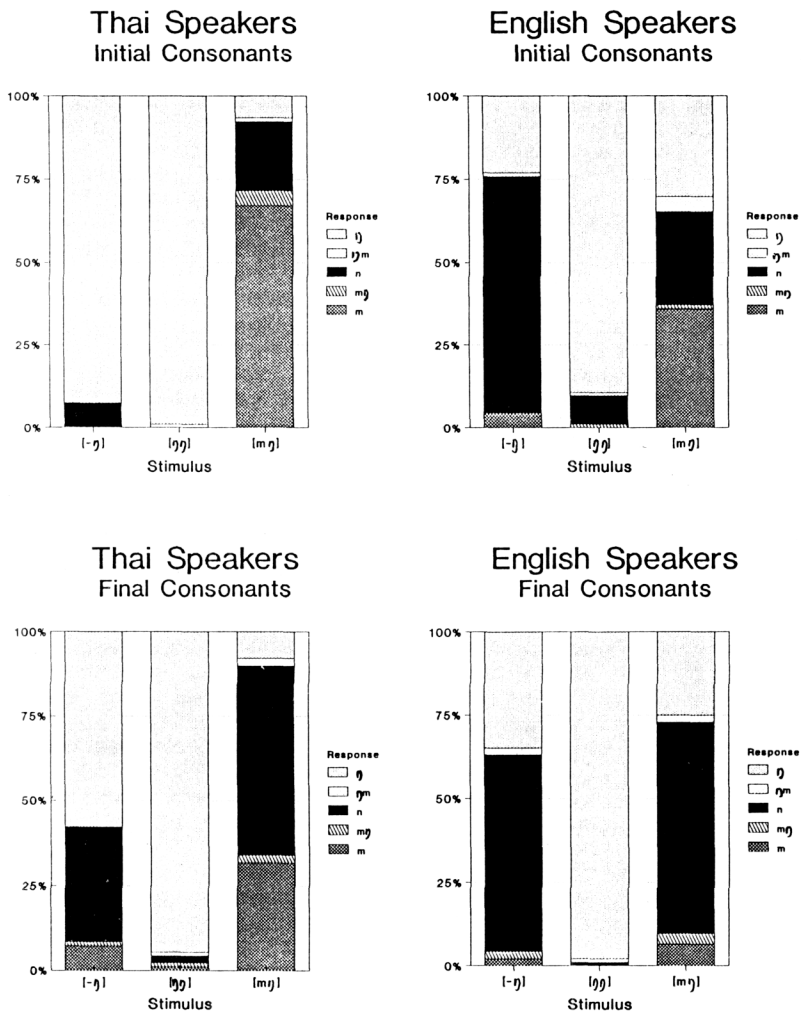
In each of the two phases the practice trials were followed by a block of 32 test trials. These consisted of four of each of the following eight stimuli: auditory only /m/, auditory only /ŋ/ visual only /m/, visual only /ŋ/, matching auditory-visual /m/, matching auditory-visual /ŋ/, combination of auditory /m/ and visual /ŋ/, and combination of auditory /ŋ/ and visual /m/. Note that all stimuli in the test phase were made up of the phonemes or visemes /m/ and /ŋ/. None of the other three practice trial stimuli (/m-ŋ/, /n/, or /ŋ-m/) were included.

At the start of the experiment, subjects were told about the task and were asked to respond on each trial as quickly and accurately as possible. If subjects made a number of slow responses during practice trials, testing was interrupted to remind them to respond as quickly as possible. At the end of the first phase (initial or final consonants phase) the experimenter explained to the subject the nature of the next phase. On completion of the experiment, subjects were informed of the true nature of the stimuli and the purpose of the experiment.

Results Only the test trials were analysed. Of particular relevance to the hypotheses were subjects' responses to the test trials on which /ŋ/ was the visual component, i.e., the visual only /ŋ/, the matching auditory-visual /ŋ/, and the auditory /m/ plus visual /ŋ/ trials. These will be labelled [-ŋ], [ŋ ŋ], and [mŋ] hereafter. The first of these, [-ŋ], provides a measure both of the visual arbiguity which occurs when the lip movements for /ŋ/ are presented alone, and of any phonological bias which may occur for English speakers when /ŋ/ is presented in the initial position. The second, [ŋ ŋ] provides a measure of phonological bias without any "contamination" from the ambiguity of presenting /ŋ/ in the visual only mode. Finally, [mŋ] is the mismatching stimulus, designed to elucidate the processes involved in the combination of auditory and visual speech elements.

Distribution of Responses In Figure 1 subjects' distribution of responses on the three trial types are shown. Interestingly, Thai speakers were able to perceive [-ŋ] veridically in the initial position, even though this viseme is thought to be ambiguous (Mills, 1987; Mills & Thiem, 1980). English speakers had many more /n/ intrusions in this initial position,

Figure 1
Distribution of Responses

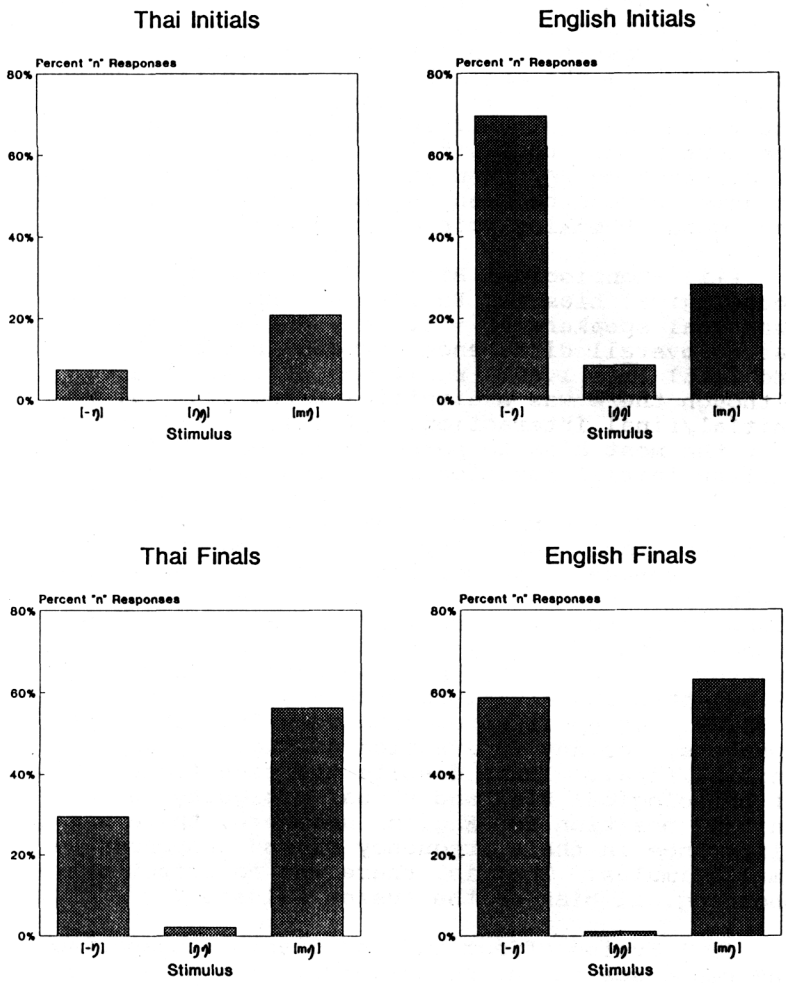


presumably due to the joint effects of phonological bias and visual ambiguity. The Thai speakers' performance was not as good in the final position, in which more /n/ intrusion errors occurred. Nevertheless their performance was still slightly better than the English subjects. This may be due to the relative infrequency of /ŋ/ in the English language - it comprises 0.92% of all English phonemes compared with 6.29% for /n/ and 2.61% for /m/ (Roberts, 1965). No similar statistics appear to be available for the Thai language, though it would seem that the overall percentage would be much greater than in English, given (a) the general use of /ŋ/ in Thai, and (b) its legality in the initial position.

In the [ŋŋ] condition most responses were correct, although in the initial position there were a number of "n" responses by English speakers. On the [mŋ] trials in all four conditions there was a distribution of responses appropriate for the auditory component ("m" responses), for the visual component ("ŋ" responses), and for the auditory-visual combination ("mŋ", "n", and "ŋ m" responses). Of the latter, "n" responses were the most prevalent. These are analysed in the next section.

Percentage "n" Responses Mean percentages of "n" responses for each of the three stimuli of interest are shown in Figure 2. Three tests based on an analysis of variance (alpha = .05 and $F_{critical}$ .05 (1,46) = 4.05) were conducted to compare initial and final consonants, and English and Thai speakers as follows. (i) The relative degree of visual ambiguity of the /ŋ/ viseme across the four conditions was tested by the percentage of "n" responses to the [-ŋ] stimulus (which measures both phonological bias and visual ambiguity) minus the percentage of "n" responses to the [ŋŋ] stimulus (which measures only phonological bias). (ii) The relative degree of phonological bias was measured by a direct comparison of all four conditions on the percentage of "n" responses to the [ŋŋ] stimulus. (iii) The effect of adding a mismatching auditory component, i.e., a fusion effect, was measured by comparison across the four conditions of the percentage of "n" responses on the mismatching [mŋ] stimulus. (iv) Finally, the [-ŋ] condition, which incorporates both visual ambiguity and phonological bias, was compared with the [mŋ] condition, in order to determine the effect of adding

Figure 2
"n" Responses - Visual /ŋ/



the /m/ auditory component onto the existing biases for visual only /ŋ/.

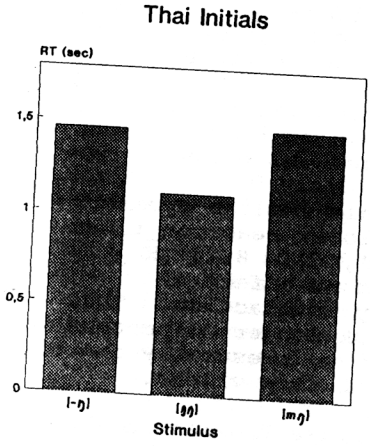
(i) Visual Ambiguity: There was an overall effect of visual ambiguity, $F(1,46) = 166.91$ which was greater for English ($\bar{X} = 59.36\%$) than Thai ($\bar{X} = 17.42\%$) speakers, $F(1,46) = 30.49$. This ambiguity was generally greater for final ($\bar{X} = 42.51\%$) than initial consonants ($\bar{X} = 34.27\%$), $F(1,46) = 5.64$, although the condition resulting in the greatest degree of visual ambiguity was the English speakers' initial consonants condition ($\bar{X} = 61.08\%$), $F(1,46) = 10.42$. Presumably, the visual ambiguity of initial /ŋ/ for English speakers is due to their lack of practice in discriminating /ŋ/ from /n/ in the initial position because, as can be seen in Figure 1, Thai subjects are very good at making this discrimination.

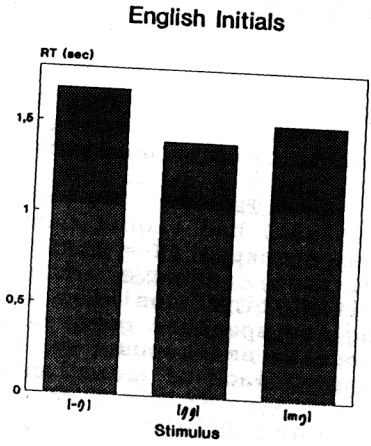
(ii) Phonological Bias: There was more phonological bias for English speakers ($\bar{X} = 9.04\%$) than Thai speakers ($\bar{X} = 1.05\%$), $F(1,46) = 4.30$. There was no overall difference between initial ($\bar{X} = 4.25\%$) and final ($\bar{X} = 1.60\%$) positions, $F(1,46) = 2.04$, although there was a significant English/Thai by initial/final interaction, $F(1,46) = 6.24$, indicating that the most bias to perceive /n/ occurred in the English initial position condition ($\bar{X} = 8.5\%$).

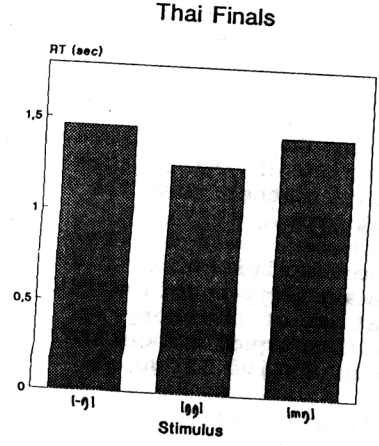
(iii) Fusion Effect: The above analyses show greatest visual ambiguity and phonological bias for English speakers' perception of /n/ in the initial position. Is this effect carried over into the mismatching fusion situation? An analysis of "n" responses for the [mŋ] condition revealed a generally greater percentage of "n" responses on final ($\bar{X} = 59.56$) than on initial ($\bar{X} = 24.50$) consonants, $F(1,46) = 34.39$, but no difference between English and Thai speakers, nor any interaction of English/Thai by initial/final. Thus, despite the significant effects of phonological bias and visual ambiguity in the initial position for English speakers, there was no difference in their frequency of "n" responses to the [mŋ] stimulus. That is, there was no effect of phonological bias on the fusion illusion.

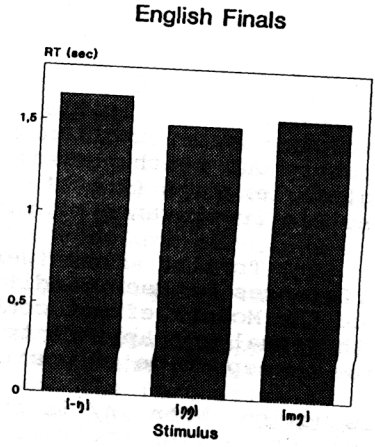
(iv) Fusion Effect and /n/ Bias: A comparison of "n" responses on the phonological bias plus visual ambiguity condition [-ŋ] and the fusion condition [mŋ] revealed a significant decrease ($\bar{X} = 41.4\%$) in "n" responses for English subjects perceiving initial

Figure 3
Reaction Times - Visual /ŋ/









consonants and a general increase for the other three conditions ($\bar{X}$ = 14.79%), F(1,46) = 16.42. That is, addition of the mismatching auditory /m/ actually resolved the uncertainty of English speakers responses to the initial [-ŋ] stimulus, while it increased the probability of "n" responses in the other three conditions. This, along with the lack of significant differences for the fusion effect, suggests that the dubbing of auditory /m/ onto visual /ŋ/ produces a consistent effect which is independent of experiential effects such as phonological bias or visual ambiguity.

Reaction Times The mean reaction times for [-ŋ], [ŋŋ] and [mŋ] stimuli are shown in Figure 3, and the results of data analyses are set out below.

[-ŋ] Trials: On [-ŋ] trials English speakers ($\bar{X}$ = 1.658sec) had significantly slower reaction times than Thai speakers ($\bar{X}$ = 1.457), F(1,46) = 8.26 ($\bar{X}_{diff}$ = 201 msec). However, there was no effect of initial/final position nor any interaction. Thus English speakers generally took longer to process the visually ambiguous [-ŋ] stimulus, presumably due to general lack of experience with this viseme.

[ŋŋ] Trials: On [ŋŋ] trials English speakers were generally slower than Thai subjects ($\bar{X}_{diff}$ = 224msec), F(1,46) = 21.71, and reactions were generally slower on final than initial conditions ($\bar{X}_{diff}$ = 240msec), F(1,46) = 7.72. This latter effect is presumably a product of the experimental procedure: as reaction times were measured from sound onset they would, other things being equal, be generally slower for final consonants. There was no interaction of English/Thai with initial/final position, indicating that the slower reaction times by English speakers are probably due to the general paucity of /ŋ/ in the English language (Roberts, 1965), rather than the specific irrelevance of initial /ŋ/.

[mŋ] Trials: There were no significant differences between conditions on [mŋ] reaction times. Thus the McGurk effect combination of auditory /m/ with visual /ŋ/ appears to require equal processing time irrespective of position or native language.

Discussion When /ŋ/ is presented in the initial position English speakers' responses are affected by the visual ambiguity of the stimulus and by the phonological bias of their native language, resulting

in a high incidence of "n" responses compared with the other three conditions (English final, Thai initial, Thai final). When the mismatching stimulus [mŋ] is presented this effect disappears - there is no longer any influence of visual ambiguity or phonological bias on the incidence of "n" responses. It is similar with the reaction times: on the [mŋ] trials there are no significant differences in reaction times, despite longer reaction times by English speakers than Thai speakers on both [-ŋ] and [ŋŋ] trials.

What can be the cause of this? It appears that the fusion effect occurs <u>independently</u> of phonological bias that speakers may have as a result of their language experience. That is, the phonetic properties of audible /m/ and visible /ŋ/ combine at a level which is uninfluenced by language-specific phonology.

Two pieces of additional information support this phonetic fusion hypothesis: (i) the differences between initial and final consonants, and (ii) the differences between responses to the [-ŋ] and the [mŋ] stimuli. With regard to the first of these, there are more "n" fusion responses in the final [mŋ] condition than in the initial [mŋ] condition for both language groups. This would appear to be due to an overall effect of visual ambiguity rather than an effect of phonological bias, because in the visual ambiguity analysis the Thai speakers, for whom both initial and final /ŋ/ are relevant, showed greater ambiguity (more "n" responses) in the final than in the initial consonant condition. On the other hand, the phonological bias analysis revealed a slightly greater bias overall on initial compared with final consonants and this was accounted for <u>wholly</u> by the English speakers. Thus it appears that subjects' responses in the [mŋ] condition are influenced by phonetically-based effects such as the greater ambiguity of visual /ŋ/ in the final position, rather than any effects of phonological bias. In short, the McGurk effect is a phonetically-based phenomenon, because it is influenced by phonetic and not phonological factors.

The second piece of information which supports the phonetic fusion hypothesis concerns the difference between responses to the [-ŋ] and the [mŋ] stimuli. There is a different relationship between the number of "n" responses on [-ŋ] and [mŋ] trials for the English initials than for the other three conditions. For these other three conditions there is an <u>increase</u>

in "n" responses from [-ŋ] to [mŋ] trials, indicating that dubbing produces less veridical responding when conflicting auditory information is added. For the English initials, however, the degree of phonological bias and visual ambiguity of initial /ŋ/ is not increased by the addition of conflicting information. Rather, this conflicting information actually serves to "resolve" the visual ambiguity and phonological bias. Thus it seems that the bias involved in perceiving /n/ for /ŋ/ in the initial position is removed when conflicting auditory information is added. Processing then proceeds at a phonetic level, obeying the same principles (final visual /ŋ/ is more ambiguous than initial visual /ŋ/) irrespective of the language background and phonological bias of the observer. It is as if the English speakers do not notice the ambiguity of initial visual /ŋ/ when other complementary phonetic information is available.

Conclusions Phonological bias does occur in speech perception but only under certain conditions: it occurs when the stimulus information is ambiguous, e.g., when a viseme such as /ŋ/, which has various possible phonetic realizations, is presented, and it also occurs as a post-access phenomenon. For example, for [-ŋ] (and the matching [ŋŋ]), the correct response is " ŋ ", a phonological oddity in the initial position in English, so English speakers are biased to respond "n".

Phonological bias does not occur when conflicting auditory and visual information are presented in a way that allows a "fusion" response which is phonologically natural in the subjects' language. For example, in the current study, on the [mŋ] trials there is a tendency for /n/ to be perceived, a phoneme which is used in English in both initial and final positions. So the results obtained in this study show that subjects' combination of auditory and visual information in the [mŋ] stimulus to perceive /n/, occurs directly at the phonetic level, and is unhindered by any post-access response bias.

The results of this study are consistent with the infant study described earlier. In that study it was found that 4-month-old infants perceive an auditory /ba/ plus visual /ga/ stimulus to be more similar to /da/ or /ða/ than to /ba/, strongly suggesting that young infants perceive the fusion effect. Young infants' have only limited phonological experience, so

their perception of the fusion effect must be due to combination of auditory and visual components at a pre-phonological level . The results of the current study with adults show that even when phonological experience is available, auditory and visual information are still combined at a pre- or sub-phonological level. This is despite the fact that under ambiguous conditions, native phonology does bias adults' perception. The results are consistent with an interactive model in which phonetic information implicit in auditory and visual sources is sufficient for perceptual resolution, but learned phonological information is incorporated in decisions when phonetic information is incomplete, ambiguous, or conflicting.

References

Burnham, D.K. (1986). Developmental loss of speech perception: Exposure to and experience with a first language. Appl. Psycholing., 7, 207-240.

Burnham, D.K., Earnshaw, L.J., & Clark, J.E. (1991). Development of categorical identification of native and non-native bilabial stops: infants, children and adults. J. Child Lang., 18, 231-260.

Dodd, B., & Burnham, D.K. (1988). Processing speech-read information. Volta Review, 90, 45-60.

Kuhl, P.K. (1978). Predispositions for the perception of speech sound categories: A species-specific phenomenon? In F.D. Minifie & L.L. Lloyd (Eds) Comm. and cognitive abilities: Early behavioural assessment. Baltimore: Univ. Park Press.

McGurk, H., & McDonald J. (1976). Hearing lips and seeing voices. Nature, 264, 746-748.

Mills, A.E. (1987) The development of phonology in the blind child. In B. Dodd & R. Campbell (Eds) Hearing by eye: The psychology of lip-reading. Hillsdale, N.J.: Erlbaum.

Mills, A.E. & Thiem, R. (1980) Auditory-visual fusions and illusions in speech perception. Linguistische Berichte, 68/80, 85-108.

Roberts, H. (1965) A statistical analysis of American English. The Hague: Mouton & Co.

Werker, J.F., McGurk, H., & Frost, P. (1987). Cross-language differences in bimodal speech perception. Paper presented at joint Exp. Psych. Ass. and Canad. Psych. Ass. Conf., Oxford, July 1-3, 1987.

Werker, J.F., & Tees, (1984). Cross-language speech perception: evidence for perceptual re-organisation during the first year of life. Infant Behav. & Develop., 27, 49-63.