WHERE DOES AUDITORY-VISUAL SPEECH INTEGRATION OCCUR? JAPANESE SPEAKERS' PERCEPTION OF THE MCGURK EFFECT AS A FUNCTION OF VOWEL ENVIRONMENT

Denis Burnham[♠], Sheila Keane[♠]

[♠]School of Psychology, University of NSW;

ABSTRACT - In the McGurk effect, when auditory [b] is dubbed onto visual [g] it is perceived as [d] or [ð]. The occurrence of this in undegraded conditions shows that human speech perceivers use visual information whenever it is available. This study uses phonetic and phonological tools to ascertain the processing level at which auditory-visual speech integration occurs. The phonetic tool is the fact that the relative frequency of [d] and [ð] fusion responses changes over vowel environment: for auditory [ba] + visual [ga] English speakers' report [ða] more often than [da], while for auditory [bi] + visual [gi], more [di] than [ði] reponses occur. The phonological tool rests on differences in phonology: while Japanese phonology contains [b], [g], and [d], it does not contain [ð]. English speakers and Japanese speakers at three levels of English proficiency, Beginner, Intermediate, Advanced, were tested on the [b] + [g] McGurk effect in [a] and [i] vowel environments. If the integration of auditory and visual speech components occurs at a phonetic level, then Japanese speakers should show appropriate shifts in frequency of [d] vs [ð] responses in the [a] vs [i] vowel conditions, despite the phonological irrelevance of [ð] in Japanese. This is indeed what occurred: despite Japanese subjects' propensity for a phonological bias towards perceiving [d] rather than [ð] in both impoverished (auditory-only [ð], and visual-only [ð]) and unimpoverished (auditory-visual [ð]) conditions, all subject groups showed a similar change in frequency of [d] and [ð] responses over the [a] vs [i] vowel conditions. The results are discussed in relation to the role of phonetic and phonological factors in auditory-visual speech integration.

When acoustic [ba] is dubbed onto the lip movements for [ga] (designated A[ba]/V[ga]), adults perceive [da] or [ða] (McGurk & McDonald, 1976). Thus visual information is not an optional adjunct used only when auditory information is degraded or absent, but an integral part of speech perception. Given this integrality, it is important for speech perception theories, to know how auditory and visual speech information are combined. Werker and Logan (1985) distinguish three types of processing in terms of the successive refining (and loss) of information from the original acoustic signal when humans process speech. In auditory processing (sensory will be used here to avoid confusion) all perceivable differences between stimuli are available; in phonetic processing only those differences which occur between phones in any language (ie, sounds with possible articulatory realizations) are available; and in phonemic processing only differences which are phonologically relevant in the perceiver's language are available. At which level does auditory-visual speech integration occur? The McGurk effect is ideal for investigating this question.

There is evidence that auditory and visual speech information is integrated beyond the sensory level, because in the auditory adaptation effect, in which adaptation to one end-point of a speech continuum shifts subjects' perception of a category boundary, adaptation to A[b$^{\omega}$]/V[g$^{\omega}$] results in auditory adaptation to [b$^{\omega}$], not auditory-visual adaptation to [d] (Roberts & Summerfield, 1981). Other studies show that integration must occur either phonetically or phonemically. For example, Green and Miller (1985) used the fact that when subjects are asked to identify sounds along a [bi]-[pi] continuum, increasing the duration of the [i] vowel increases the number of [b] responses. This same effect was found even when the vowel duration was specified by lip movements, with the acoustic vowel duration remaining constant. Thus phonetic features interact across modalities, and the McGurk effect appears to occur in a space where at least phonetic features are available.

Thus auditory-visual speech integration is not simply a sensory effect and involves either language-general (phonetic) or language-specific (phonemic) processes. We hypothesise that auditory-visual speech integration occurs without reference to the ambient language environment, ie, that it is a phonetic phenomenon. If so, then two predictions can be made. First, the McGurk effect should occur pre-phonologically, in infancy, before phonemic processing is strongly evident. Second, it should occur sub-phonologically, cross-linguistically, transcending the phonological constraints of particular languages.

Recent studies suggest that the McGurk effect does in fact occur in infancy. Johnson, Rosenblum and Schmuckler (1995) habituauted 5-month-old infants' visual fixation to auditory-visual [va], and then gave three test trials: the original A[va]V[va], A[ba]V[va] (perceived by adults as "va"), and A[da]V[va] (perceived by adults as "da"). The rate of re-habituation was slower in A[da]V[va] than A[va]V[va] and A[ba]V[va], the latter two being equivalent. Thus, as is the case for adults, A[ba]V[va] was perceptually equivalent to

A[va]V[va] for 5-month-olds, while A[da]V[va], which gives rise to a different percept in adults, "da", was not. These results are interesting however, they only show that infants take account of both auditory and visual components in dubbed AV stimuli, because the version of the McGurk effect used did not elicit a *fused* response, ie, it did not involve the emergent perception of a third phone which was neither the auditory nor the visual component.

Burnham and Dodd (1996; Burnham, 1992) tested 4 1/2 month-old infants for their fused percept of [d] or [ð] in the traditional A[b]V[g] McGurk effect. In habituation the experimental group were presented with A[ba]V[ga] and the control group with matching A[ba]V[ba]. In the three test trials both groups were presented with Aud-only [ba], [da] or [ða]. If experimental group infants had perceived a [da] or [ða] fusion then one of these sounds plus [ba] would appear novel to them, while if their perception was auditorally governed then both [da] and [ða] should appear novel, but not [ba]. Control group infants' test responses should, of course, be governed by the auditory percept, [ba]. Fusion and auditory hypotheses were tested for each group based on "percentage range accounted for" auditory and fusion scores derived from relative fixation times on the three test trials. The control group had higher auditory than fusion scores, while the experimental group had higher fusion than auditory scores. Thus 4 1/2-month-old infants perceive a fused percept when presented with the A[b]V[g] McGurk. While these results suggest that auditory-visual speech integration occurs at a phonetic level of processing, it is not clear whether infants at this age are free from the constraints and influences of the ambient phonology. One way to investigate this further is to conduct further research with younger infants. Another is to seek convergent evidence on phonetic processes in auditory-visual speech perception from cross-linguistic studies.

In the first study of this nature, Werker, Frost and McGurk (1992) paired auditory [ba] with visual [ba], [va], [ða], [da], [ga], or [ga] and asked ELE subjects and native French subjects with varying levels of English proficiency for perceptual identifications. All these phones are relevant except for [ð]. There were more [ða] identifications for A[ba]V[ða] by English than by French subjects, who tended to substitute [da] or [t^ha]. However, the frequency of [ða] identifications increased as a function of English language experience, suggesting some influence of phonology. Two explanations are possible: French subjects may have, through their language experience, a *phonological bias* against perceiving [ð]. However, it could equally be a post-perceptual labelling effect: French subjects may perceive [ð] equally 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. ensured that the French subjects could produce [ð] and also found no difference in the incidence of written and spoken "d" or "t" responses to acoustic-only [ð] presentations, it is nevertheless possible that there was a response bias against a phone not used in the perceivers' native phonology and orthography.

Burnham (Burnham 1992; Burnham & Dodd, 1996) conducted a study to overcome the confounding effect of response bias. The fact that [ŋ], as in sing, is used in both word initial and final positions in Thai but only in the final position in English was exploited to investigate the effect of phonology on auditory-visual speech perception. Thai and English adults were presented with A[m]V[ŋ] either in syllable-final or syllable-initial position. This overcomes the response bias problem because the non-native phoneme is a *component* of the McGurk effect, rather than the *response*. English subjects' perception in the initial position is the condition of interest. They would be be expected to respond "n" more often in the initial than in the final position for Vis-only [ŋ] and for matching AV [ŋ]. The results showed that English speakers had a greater bias to respond "n" in the initial position for both visual-only [ŋ] and matching auditory-visual [ŋ]. However when "n" responses to A[m]V[ŋ] were investigated, it was found that there was no difference between English and Thai subjects, nor any interaction of English/Thai x initial/final. Thus, despite the greater phonological bias for initial AV [ŋ] and visual ambiguity of initial Vis-only [ŋ] for English speakers, their frequency of "n" responses to A[m]V[ŋ] was not significantly greater than that of the Thais. The study shows that there was no effect of phonological bias on the McGurk effect: the incidence of "n" responses for the McGurk are equivalent, irrespective of position or native language.

This conclusion would be strengthened if the same result were found in other situations, and when the phonogically-irrelevant phone was not only phonotactically-irrelevant (as in the case of the non-initial use of [ŋ] in English), but also phonemically-irrelevant. In addition, it would be good if this were to be shown when the phonemically-irrelevant phone occurred as the emergent perception rather than as one of the components. Both of these conditions hold in the use of A[b]V[g] (perceived as [ð] or [d]), (Werker, Frost & McGurk, 1992), but as pointed out earlier, Burnham suggests that possible problems of response bias were not overcome in this study (Burnham 1992; Burnham & Dodd, 1996).

THE EXPERIMENT

The current study uses the A[b]V[g] stimulus, and a well-established phonetic effect to test the occurrence of the McGurk effect in English speakers, and Japanese speakers at three levels of English proficiency (Beginners, Intermediate, and Avanced). In Japanese [d] is phonologically relevant, while [ð] is not. Green (1996) has shown that the relative distribution of "d' and "th" responses to the traditional A[b]V[g] McGurk changes when moving from an [a] to [i] vowel context: for English speaking listeners the incidence of "d" responses increases and "th" responses decreases in an [i] vowel context. If the McGurk effect occurs at a phonetic level, then this phonetic effect should be apparent in both English and Japanese speakers, irrespective of their experience with English and therefore with [ð].

METHOD

Subjects

Forty-eight native Japanese adults and 16 native English adults were tested. The Japanese groups had Beginner, Intermediate or Advanced level of English proficiency (English Language Intensive Courses for Overseas Students), and 16 were tested in each group. In the Beginner group, there were 11 females and five males, and they had been in Australia for six weeks on average (range 1 week - 3.5 months) prior to testing. There were seven females and nine males in the Intermediate group, and they had spent three months on average (range 2 weeks - 12 months) in Australia. Japanese speakers for the Advanced group were required to have worked in an English-speaking environment in Australia/New Zealand for at least one year but not more than ten. Twelve females and four males, comprised this group, and on average, these subjects had lived in Australia for three years (range 1.1 years - 8 years). For the English Language control group, ten female and six male Australians were recruited from the first-year psychology student pool at the University of NSW.

Stimulus Construction

The stimuli consisted of Aud-only, Vis-only and matching AV presentations of the syllables [bV], [gV], [dV], [ðV], [bgV] and [gbV] and mismatching presentations of Aud[bV]-Vis[gV] and Aud[gV]-Vis[bV], where V stands for the vowel context, [aᵅ] or [iᵅ]. The stimuli were prepared by video-recording the head and shoulders of a female native (Australian) English speaker, who fixated the camera as she presented each syllable once every four seconds for twelve repetitions. When articulating the consonant clusters, she was asked to insert the schwa vowel, /«/, between consonants, e.g., [b«gaᵅ] and and [g«baᵅ]. The speaker was also recorded with a motionless face for four seconds. This was necessary for the later construction of the auditory-only stimuli.

Visual Components: Four visual exemplars of each syllable were selected and edited to provide the visual component for the Aud-only, Vis-only and AV stimuli. The stimuli for each vowel condition were edited onto separate video-tapes according to a previously determined random order. The order was restricted to ensure that not more than two stimuli of the same type e.g. Aud-only or Vis-only, or of the same sound e.g. [b], [d] etc., appeared consecutively. Each visual stimulus lasted for four seconds with one second of black background intervening between each trial. For the Vis-only and AV trials this consisted of (i) one second of a motionless face, (ii) approximately one second of articulation and (iii) two more seconds of neutral expression. For the Aud-only trials, the speaker's motionless face was presented for four seconds.

Auditory Components: Auditory syllables, all approximately equal duration, were selected from the original video-tape and digitised using signal processing software (Kay CSL 4500 package). Auditory stimulus durations were matched within and between syllables for [bV] and [gV]. Three exemplars of each of the six syllables [bV], [gV], [dV], [ðV], [bgV] and [gbV] were used in each of the two vowel contexts [aᵅ] and [iᵅ]. These were dubbed onto the visual stimuli in real time, this being controlled by means of a locally-produced computer program, MAKEFUS. For the Aud-only trials, speech sounds were presented without any corresponding lip movements.

The AV matching and mismatching stimuli were created by combining the auditory and visual components. The original sound from the video-recording was conveyed through one audio channel on the video-recorder, to the voice-key box which activated a digital input to the computer, and thence the appropriate digitised sound from disk. The original stimulus itself was inaudible to the subject. On the matching trials, the auditory and visual components corresponded, and were presented simultaneously. The mismatching trials were of two types i.e. Aud[bV] was dubbed onto Vis[gV] or Aud[gV] was dubbed onto Vis[bV]. To ensure conformity across conditions all sounds for both the matching and mismatching AV trials were dubbed.

On the Vis-only trials, the auditory stimulus from the video-tape cued the computer to play 'silence' from disk and thus, just the lip movements without any accompanying sound was presented. For the Aud-only trials, in

which a motionless face was presented, an auditory stimulus was triggered from disk by a tone pre-recorded in the appropriate place on the B audio channel. As well as activating sound files from disk, both the tone inputs and auditory speech signals triggered a computer clock to begin. This clock was stopped by subjects' responses, so subjects' RTs were recorded from the onset of sound for each trial.

Stimulus Trials

For each vowel condition there were eighteen practice trials, of one presentation of each of the six syllables [bV], [gV], [dV], [ɸV], [bgV] and [gbV], in each of the three different modes i.e. Aud-only, Vis-only, and AV. The purpose of these trials was to give subjects practice at using the alternative response keys and at meeting the trial time limit. The results of the practice trials were used in the initial analysis to eliminate the test data of those subjects who responded too slowly and/or too inaccurately.

For each vowel condition, there were two 32-trial test blocks i.e., a total of 64 trials. Each block consisted of exactly the same trial types however, the presentation order of the trials varied between blocks. The sequence of the test blocks was counterbalanced between subjects. In each block the subjects were presented with two Aud-only, two Vis-only and two AV matching presentations of each of the syllables [bV], [gV], [dV] and [ɸV], and four each of the mismatching Aud[bV]-Vis[gV], and mismatching Aud[gV]-Vis[bV].

Apparatus and Procedure

Subjects were tested individually in a small sound-attenuated room. They sat in front of a small television monitor connected to a video-recorder in the control room. A response pad was placed directly in front of the monitor. This had a central 'ready' key surrounded by six response buttons. The buttons were arranged in a semicircle, equidistant from the 'ready' key and labelled as 'b', 'g', 'd', 'th', 'bg' and 'gb' in all conditions for all subjects. The particular configuration used for the response keys was the one found, in a short pilot study, to provide the optimum conditions for responding. The six response keys and the 'ready' key of the response pad, served as seven separate inputs to the computer in the control room. A reward light output from the computer was attached to the left side of the monitor. This flashed only during the practice trials, and only when subjects responded correctly. An error buzzer output from the computer was placed in the testing chamber and sounded to inform the subjects and experimenter of any failure to respond according to the appropriate procedure, e.g. if response durations were too long.

In the control room an IBM-compatible 386 computer was used to control the presentation of stimuli from the video-recorder and to record subjects' responses. The original audio output from the video served as digital input to the computer, via a voice-key box. These inputs were then used to generate the appropriate sounds from disk which were in turn routed through D-A and filter boards to the speaker on the monitor in the test chamber. The auditory stimuli therefore corresponded temporally with the visual stimuli which were simultaneously presented to the subject in the next room. Subjects' responses on each trial were recorded on disk.

The order of conditions was counterbalanced across groups. In each condition there was one block of practice trials followed by two test trial blocks. On any particular trial the subject was told to press the 'ready' key to start each trial. As each stimulus was presented the subject was required to respond as quickly and accurately as possible, by pressing the response button which "best matched the consonant sound the speaker used". After each response was made the subject returned their finger to the 'ready' key as a signal to the computer that they were ready for the next trial. If the subject failed to respond within the maximum time limit of 3.5secs. an error buzzer sounded to inform both the subject and the experimenter of a null trial. If the subject took their finger off the 'ready' button prior to the onset of the sound the error buzzer signalled that this had occurred and again, a null trial was recorded. Testing duration was approximately 30 minutes.

RESULTS

The results for speakers' "th" responses to AV, A-only, and V-only [ɸ] are shown in Figure 1. As can be seen Japanese speakers made more errors than Australian English speakers on AV, $F(1,60) = 11.16$, A-only, $F(1,60) = 24.26$, and V-only, $F(1,60) = 7.39$. However, their performance improved linearly as a function of English language experience in AV, $F(1,60) = 8.03$, A-only, $F(1,60) = 10.04$, but for V-only only the quadratic trend was significant, $F(1,60) = 7.43$. In addition, Figure 1 shows the number of cluster responses, "bg" or "gb", to A[g]V[b], the opposite pairing to that used in the McGurk effect. As can be seen, English speakers gave more cluster responses than did the Japanese speakers combined, $F(1,60) = 6.13$. This result can be attributed to the lack of consonant clusters in Japanese phonology.

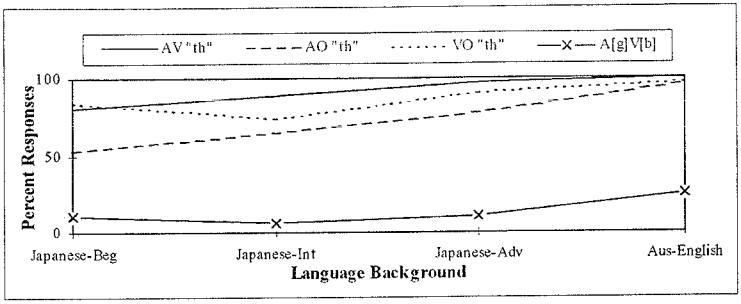

Figure 1. Percent "th" Responses on AV, A-only, and V-only presentations of [ð], and of cluster responses on A[g]V[b] trials.

So Japanese speakers show appropriate biases in accord with their native phonology. What happens when the A[b]V[g] McGurk is presented? Are they affected by the phonetic vowel context despite their bias towards perceiving "d" for [ð]. Analysis revealed that the Japanese gave more "th" responses than "d" responses compared to the English speakers, $F91,60) = 31.31$, again indicating Japanese speakers' phonological bias. In addition, there was, as expected, a effect for vowel type x "d"/"th" responses, and this was statistically equivalent for all four groups. That is there was *no* significant interaction of this effect with language group, even though a number of post-hoc tests were conducted. As can be seen in Figure 2, the vowel context effect occurred for all four language groups, although somewaht differently in some. For example, the Intermediate group increased both their "d" and "th" responses from the [a] to the[i] vowel context, but in accord with the general prediction, the increase for "d" was greater than that for "th". These results show that auditory and visual speech information are initially integrated at a phonetic level of processing free from phonological constraints. Any phonological constraints that influence subjects' responses then occur post-phonetically on the resultant auditory-visual speech percept.

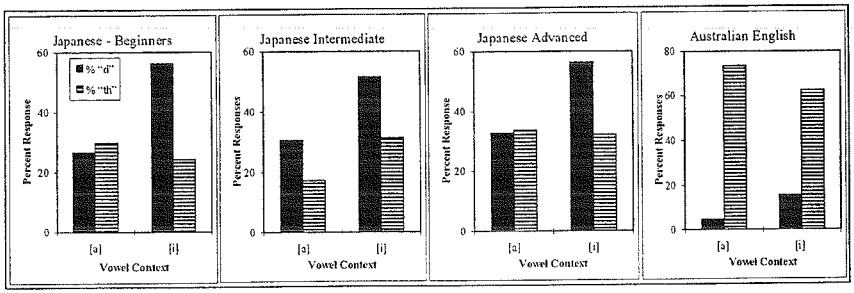

Figure 2. Effect of Vowel Context on the Incidence of "d" and "th" Responses in the McGurk Effect.

CONCLUSIONS

These results confirm those of Burnham and Dodd (1995) who found that phonological constraints do not influence the incidence of the McGurk effect. English speakers there had a bias to perceive [n] for [ŋ] in the initial position, but despite this, the incidence of [n] fusions in the A[m]V[ŋ] McGurk was equivalent for Thai and English speakers. Here, Japanese speakers had a phonological bias to perceive "d" in situations where "th" was correct. Nevertheless, their distribution of "d" and "th" responses in [a] and [I] vowel environments followed the same pattern as that for English speakers. Together these results provide strong evidence that auditory-visual integration of speech information occurs *sub*-phonologically. This does not deny that there may be post-access effects of the speakers' phonology on their responses, but the initial integration of information occurs at a phonetic level of processing.

REFERENCES

Burnham, D. K. (1992) Processing auditory-visual speech in infancy and across phonologies. *International Journal of Psychology, 27*, 59.

Burnham, D., & Dodd, B. (1996) Auditory-visual speech perception as a direct process: The McGurk effect in infants and across languages. In D. Stork and M. Hennecke (Eds.) *Speechreading by humans and machines.* Berlin: Springer-Verlag.

Green, K. (1996) The use of auditory and visual information in phonetic perception. In D.G. Stork & M.E. Hennecke (Eds.) *Speechreading by humans and machines.* Berlin: Springer-Verlag.

Green, K.P., & Miller, J.L. (1985) On the role of visual rate information in phonetic perception. *Perception and Psychophysics, 38*, 269-276.

Johnson, J.A., Rosenblum, L.D. & Schmuckler, M.A. (1995) The McGurk effect in infants. *Journal of Acoustical Society of America 97*, 2aSC7. 3286.

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

Werker, J.F., Frost, P., & McGurk, H. (1992). Cross-language influences on bimodal speech perception. *Canadian Journal of Psychology, 46*, 551-568.

Werker, J.F., & Logan, J.S. (1985) Cross-language evidence for three factors in speech perception. *Perception and Psychophysics, 37*, 35-44.