Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

Audio-visual speech perception in noise by first and second language listeners

Michael Fitzpatrick (1), Jeesun Kim (1)

(1) MARCS Auditory Laboratories, University of Western Sydney, Australia

PACS: 43.71.Hw, 43.72.Dv

ABSTRACT

Second language (L2) listeners' auditory speech perception is more vulnerable to noise than that of first language (L1) listeners. Impoverished auditory perception may cause L2 listeners to rely more on visual speech cues when perceiving speech in noise. The present study examined whether L1 and L2 perceivers might differ in their use of visual speech cues. In the experiment English-Spanish and Spanish-English bilingual participants were tested in a phoneme identification task across 16 English and 16 Spanish consonants (in the context of VCV syllables) that were presented in auditory-only, visual-only and auditory-visual conditions, with or without background 'babble' noise. The results showed that overall, L1 perceivers outperformed L2 perceivers across all conditions, and both groups improved in auditory-visual compared to auditory-only conditions. L2 listeners' performance showed a greater drop from in-clear to in-noise conditions compared to L1 listeners. Despite the discrepancy between L1 and L2 listeners in performance, the relative degree of improvement in auditory-visual compared to auditory-only conditions was the same for both L1 and L2 listeners. Further, auditory-visual integration efficiency measures showed no significant difference between the L1 and L2 listener groups. These results suggest that L1 and L2 users give similar weight to visual cues in speech perception and indicate that L2 listeners' vulnerability in perceiving acoustic speech cues in noise is not compensated for by better use of visual speech cues.

INTRODUCTION

Infants are sensitive to phonological contrasts from almost any language; however, after about 6-10 months of exposure to a language the ability to perceive non-native contrasts shows a marked decline [1, 2]. A consequence of this phonetic organisation in infancy and childhood is the difficulty of perceiving speech in L2. That is, although affected by many factors such as the specific L1 already acquired, age of learning and the frequency of use of the L2, it appears that perceiving speech in an L2 is never as proficient as in L1 [e.g., 3, 4].

L2 speech perception in noise provides a clear cut example of the difficulty that attends L2 listening. Here, the typical finding is that L2 speech perception in noise is significantly worse than that of L1 listeners [e.g., 5, 6, 7, 8]. Even bilingual participants, who can demonstrate almost equal speech perception to native listeners in clear speech conditions, have been shown to perform significantly worse than native monolingual listeners when speech is masked by noise [9].

This fragility of L2 perception in noise has been recently demonstrated at the level of phoneme identification. For example, Garcia Lecumberri and Cooke (2006) tested American English and Spanish participants' perception of American English consonants across a range of different noise maskers. They found that the noise affected the L2 listeners significantly more than the L1 listeners; the L1 and L2 participants performing comparably well in clear conditions (the two groups differed by 7% points), but the L2 listeners performing significantly worse than L1 listeners in noise conditions (the two groups differed to a maximum of 18% for 'babble' speech at 0dB signal-to-noise ratio (SNR)). In a subsequent study conducted by Cutler, Garcia Lecumberri and Cooke (2008) using the same stimuli materials but with a Dutch L2 group, the same L2 vulnerability to noise result was found.

One possible explanation for the relative vulnerability of L2 compared to L1 listening in noise (particularly in regard to phoneme identification), is that through extensive experience with perceiving L1 phonemes (which includes experience in listening to their L1 speech in adverse conditions), listeners develop multiple cues for speech perception [7, 8]. As such, although masking noise may degrade some cues for auditory speech identification, sufficient cues remain that can be exploited by L1 listeners to facilitate speech perception and overcome noise masking effects. In contrast, L2 listeners with more limited experience might either not be sensitive to mulitple cues, or have the ability to exploit such cues, thus, L2 speech perception in noise will be poorer relative to L1 [7, 8].

The vast majority of research into L2 speech perception (especially L2 speech perception in noise) has focussed primarily on the auditory domain [e.g., 10, 11]. It is, however, well established that speech is a multimodal phenomenon in the sense that the visual speech (e.g., the perception of the motion of the lip, teeth, tongue and peri-oral facial features that occur during speech production) significantly influences the perception of speech [e.g., 12, 13, 14, 15, 16]. Furthermore, when the auditory signal is degraded the cues provided by visual speech (which are unaffected by auditory masking) can be used to strengthen and disambiguate the auditory signal and thus substantially improve speech perception in noise [e.g., 17, 18, 19, 20].

It is unclear therefore whether L2 listeners show the same vulnerability for L2 phoneme identification in noise with AV presentations, as they do in auditory alone presentations. On the one hand, visual speech information may compensate for the L2 listeners' difficulty in auditory speech perception in noise [e.g., 21]. That is, the incorporation of visual speech may provide the necessary multiple redundant cues to make L2 speech perception in noise more robust, and alleviate the L2 relative to L1 speech in noise vulnerability.

On the other hand, however, the inclusion of visual speech may not necessarily translate into improved L2 speech perception in noise [e.g. 22]. The ability to extract and use L2 visual speech cues (similarly to L2 auditory cues) may require extensive exposure and experience [23]. For example, L2 listeners might not be as adept at perceiving L2 visual speech, or they might be less efficient at integrating the available auditory and visual speech cues in L2 AV speech [e.g., 24, 25]. In this latter case, although L2 listeners may adequately perceive L2 visual speech, they may not be able to use it as efficiently as they do L1 visual speech. In either case, visual speech would not provide the same degree of benefit for L2 listeners as it does for L1 listeners and so L2 visual speech would not compensate for L2 auditory speech perception.

In order to investigate the effect of noise on L2 auditory speech perception and determine how this might be affected by the provision of visual speech, the current study examined L1 and L2 listeners' perception of consonants, in clear and in noise conditions, in auditory only (AO), visual only (VO) and auditory visual (AV) conditions. By examining these presentation conditions, the study both replicates and extends existing work on L2 speech perception in noise.

In the current study several design considerations were taken into account. Firstly, it is important to note that any visual influence on L2 speech perception can potentially be affected by variables other than the listener's AV integration efficiency. Such factors include the relationship between the L1 and the L2 [e.g., 10] language-specific weighting to visual cues [e.g. 25], or cultural factors [e.g., 24]. Given this, it is necessary to select L1 and L2 languages where a similar influence of visual speech has been demonstrated. In this regard, the current study selected English and Spanish as similar AV weighting has been demonstrated for each language when tested with L1 participants [26]. Further, an attempt was made to control for the degree of L2 experience by recruiting English-Spanish and Spanish-English beginning bilingual participants. In this way, the two different language groups can be collapsed for the L1 and L2 comparison and any L1-L2 difference due simply to L1 and L2 experience should be minimised.

Second, a shortcoming of many investigations of visual influence in speech perception has been that they have tended to use extremely limited sets of stimuli (i.e., the majority of research into L2 AV integration have used the McGurk paradigm, which usually consists of /ba/ /da/ /ga/). As some of the difficulties experienced by L2 listeners in auditory and AV speech perception may be due to the different L2 consonants, it is important to use a wide range of stimuli in investigating L2 speech perception. Further, recent studies [e.g., 27] have demonstrated that individual variation can be a factor in determining the amount of visual speech that listeners perceive from talkers. As such, the current study examined L2 speech integration using a range of consonants, and a number of different talkers.

METHOD

Materials

Two stimuli sets of 16 Australian English and Spanish consonants were selected. The English stimuli were /b, ʧ, d, f, g, k, l, m, n, p, r, s, ʃ, θ, v, z/ and the Spanish stimuli were /b, ʧ, d, f, g, k, l, m, n, ɲ, p, ɾ, r, s, t, ʎ /. Each of the consonants for both language sets were embedded in a vowel-consonant-vowel context (the vowel used being /a/, e.g., /aba/). This context was selected as it provides a more consistent environment for consonant identification compared to (say) high back and front vowels [6], and a constant consonant context across stimuli helps minimize coarticulatory differences.

These consonants were selected such that an adequate range of speech components (voice, manner, place of articulation) were represented within each language. Some of these consonants existed in both languages (e.g., English and Spanish /k, m/) and some were specific to only one of the languages (e.g., English /v/, Spanish /ɾ /).

To create the stimuli for the experiment, 6 male speakers were recruited: Three Australian English aged between 21 and 27 years (M = 24 yrs) and three South American Spanish speakers aged between 24 and 34 years (M = 29 yrs) to record the English and Spanish stimuli respectively. The stimulus recordings were conducted in a sound proof booth at MARCS Auditory Laboratories, University of Western Sydney. The speaker's face, neck and shoulders were recorded against a light blue background and were illuminated with a key and fill light. All recordings were made using a Cannon XL-1 DV camcorder and the audio was recorded from a Bruel and Kjaer type 4165 microphone to the camcorder. Multiple tokens were recorded, among which two tokens from each talker were selected for use in the testing session. Selection of the files to be used in the test session was made on the basis of clarity of pronunciation, similarity of facial movements, and of similarity of time durations to allow for consistent start and finish times across the items. Two examples of each consonant for the two languages were also recorded from a female Spanish-English bilingual speaker to be used as practice trials in the experiment.

The video files were transferred to PC and were compressed and scaled to 250*300 pixels, 25 f/s, and audio sampling rate of 44.1 khz using VirtualDUB software. The files were edited such that the beginning and end of each token showed a neutral expression with the mouth closed. The average duration of each recording file was 1.2 seconds (standard deviation = 0.18 seconds).

Three types of stimuli were produced from the recordings: Audio only (AO), visual only (VO) and auditory visual (AV) stimuli. The AO and VO items were created by separating the audio and video streams using VirtualDub software. The audio streams were saved as wav files and were normalised with PRAAT such that their peak amplitude was 65dB. The silent VO tokens were saved as avi files. The AV items were created by realigning the normalised auditory items with the visual items.

In the experiment there were three levels of auditory noise conditions for each of the AO and AV token sets for the two languages: Clear, 0 dB, and -8 dB SNRs. The noise consisted of 6 talker babble English or Spanish speech for the respective stimuli languages. The English babble consisted of 3 female and 3 male native Australian English talkers, with the Spanish babble consisting of 3 female and 3 male South American Spanish talkers. Both groups of talkers were re-

corded saying several nonsense sentences in their L1 directly onto a PC, using a Bruel and Kjaer type 4165 microphone. The recordings were equalised for RMS amplitude and were summed together into single recordings (25 seconds in duration for English; 25 seconds for Spanish) using Audacity software – the result being two sets of recordings of multiple talkers speaking at once, without any individual words being distinguishable from the noise. Both the English and Spanish babble noise recordings were combined with their respective speech stimuli (i.e. the English babble with English stimuli, and the Spanish babble with Spanish stimuli) at two SNRs: 0 dB and -8 dB (stimuli amplitude/babble amplitude) using Matlab software. The Matlab software selected a random segment of exactly 0.5 seconds longer than the speech token, so that no token had exactly the same babble noise masking it.

In sum, the experiment included three presentation conditions: AO, VO and AV. Within each of the AO and AV presentation conditions, there were three noise levels of auditory stimuli: Clear, 0 dB SNR, and -8 dB SNR. Each of these 7 (3 * AO, 3 * AV and 1 * VO) conditions consisted of 96 different tokens (16 syllables * 3 talkers * 2 tokens). This was repeated across two languages (English and Spanish) so the entire test was comprised of 1344 tokens (672 English tokens, and 672 Spanish tokens).

Participants

Twenty Three early bilingual participants were recruited for this study: 11 native Australian English Listeners (L2 Spanish) and 9 native South American Spanish listeners (L2 English). The Australian participants were recruited from a Spanish Language school in Sydney Australia. Their ages ranged from 20-45 ($\underline{M}$ = 27 yrs). The Spanish speaking participants were recruited from an English Language school for new migrants to Australia. In order to control for differences in dialect between Latin-American Spanish and Castilian Spanish, participants were restricted to native to South America. Their ages ranged from 20-48 years ($\underline{M}$ = 34 yrs). All participants reported to speak Spanish as the predominant language they currently use, and as the only language spoken at home.

All of the participants in the study were volunteers. All of the participants reported normal or corrected to normal vision and hearing.

Procedure

Both verbal and written instructions were given to each participant before running the experiment. In the experiment, participants were told that they would hear a series of English or Spanish speech stimuli (e.g., /aba/, /aga/, etc) presented one by one, and their task was to identify what consonant was included in each item (e.g., /b/ in /aba/). The participants were instructed to respond by clicking (with the mouse) one of the response items in a 4x4 grid representing the 16 consonants for the language. The 16 responses were positioned in alphabetical order with respect to their orthography in order to minimise the time taken to find the desired response.

For all participants, the two sets of language stimuli were presented in blocks, and the presentation order of these blocks was counterbalanced across participants (i.e., each participant either completed all of the English stimuli items first, followed by all of the Spanish items or vice versa). Within each of the two language stimuli, the presentation order was always AO or VO first (the order was counterbalanced across participants), followed by the AV condition. Within the AO and AV conditions the two noise conditions were always presented before the clear condition to reduce any potential learning effects. The presentation order of noise stimuli was also quasi-randomised to reduce errors due to suddenly increasing or decreasing the SNR of stimuli played after one another.

In order to familiarise participants with the task and the location of the responses on the response grid, at the beginning of each language block, the participants completed a practice session that consisted of the 32 sample stimuli (2 * each consonant) taken from a female Spanish-English bilingual talker. The practice stimuli were always presented in an AO condition, and with no noise.

For all participants, the test was run individually on a laptop PC in a sound attenuated booth. Test presentation was controlled by the DMDX software program [28]. Audio stimuli were presented through Sennheiser HD 25 headphones. The visual stimuli were presented in the centre of the screen. The entire experiment took around 2.5 hours to complete which included instructions, debriefing, and a 10 minute break between language stimuli conditions. The participants also had opportunity to break between each of the presentation conditions.

RESULTS AND DISCUSSION

The aim of the current set of analyses was to examine differences between L1 and L2 speech perception as a function of background noise level when speech was presented in auditory only (AO), visual only (VO) and auditory visual (AV) conditions. In this regard, the data of the two L1 groups and the two L2 groups were each collapsed across language so that when the L1 and L2 comparisons were made, any difference due simply to differences between English and Spanish should be minimised. Reference to an 'L1' or 'L2' group in the following refers to these combined participant groups unless otherwise specified.

For each of the AO, VO and AV conditions, the results were averaged across the three talkers in each condition. Three performance measures were calculated: firstly, overall percentage correct scores were calculated by averaging only the correct responses for each participant; secondly, confusion matrices were constructed (based on the both the correct and incorrect responses for each participant) and, using the pattern of responses detailed in the confusion matricies, transmitted information (TI) [29] rates for the articulatory-acoustic features of voicing, manner and place of articulation were calculated using the Sequential INFormation Analysis (SINFA) algorithm [30]; thirdly, in order to measure whether L1 and L2 listeners differed in the efficiency by which they integrated the available auditory and visual information, estimates of L1 and L2 listeners' Integration efficiency (IE) were calculated using PROB model (detailed further below) for the AV conditions [31]. To ease discussion of the results, the percentage correct, TI and IE scores will be presented separately.

Percentage Correct scores:

Auditory Only (AO) conditions

The percentage correct results for the L1 and L2 groups in the AO conditions as a function of noise level are presented in figure 1. As can be seen, L1 listeners were consistently more accurate than L2 listeners, and the performance of both L1 and L2 participants deteriorated considerably as the noise levels increased from clear to 0dB SNR and -8dB SNR.

The AO percentage correct scores were analysed with a mixed repeated measures ANOVA with Participant Language (English L1, Spanish L1 participants) as a between-

participant factor and Noise level (Clear; 0dB SNR; -8dB SNR) and Listener group (L1, L2 listeners) as within-group factors. The aim was to ascertain whether the identification of L2 phonemes was significantly worse in noise conditions than was the identification of L1 phonemes. The results showed the expected main effects across the L1 and L2 groups, $F(1, 18) = 62.96$, $p < .001$, partial $\eta^2 = .79$, in which participants identified significantly more L1 consonants correctly than L2 consonants and a main effect of Noise level $F(2, 36) = 989.15$, $p < .001$, partial $\eta^2 = .98$, in which the perception of both L1 and L2 consonants deteriorated as a function of increasing noise level. The between-subjects main effect comparing the English and Spanish participants was not significant ($F(1, 18) = 2.58$, $p = .126$, partial $\eta^2 = .13$), indicating that averaged across the noise levels and listener groups, both English and Spanish participants performed at similar levels of proficiency in the AO conditions.

Importantly however, the interaction effect for Noise by Listener was significant $F(2, 36) = 8.53$, $p < .001$, partial $\eta^2 = .32$, indicating L2 vulnerability to noise as reported in previous studies [e.g., In planned pairwise comparisons (with a Bonferroni adjustment made for multiple comparisons) the Listener group by Noise level interaction was significant between clear and -8dB SNR noise conditions, $F(1, 18) = 19.69$, $p < .001$, partial $\eta^2 = .522$, but not between the clear and 0dB SNR conditions, $F(1,18) = 5.89$, $p = .026$, partial $\eta^2 = .247$ or between the 0dB SNR and -8dB SNR conditions, $F(1, 18) = 2.21$, $p = .154$, partial $\eta^2 = .109$. However, the overall trend was clear: the difference between L1 and L2 phoneme perception was significantly larger in noise conditions than it was in clear; the mean L1/L2 difference increasing from 8% in clear to 13% in 0dB SNR, and 16% at -8dB SNR.

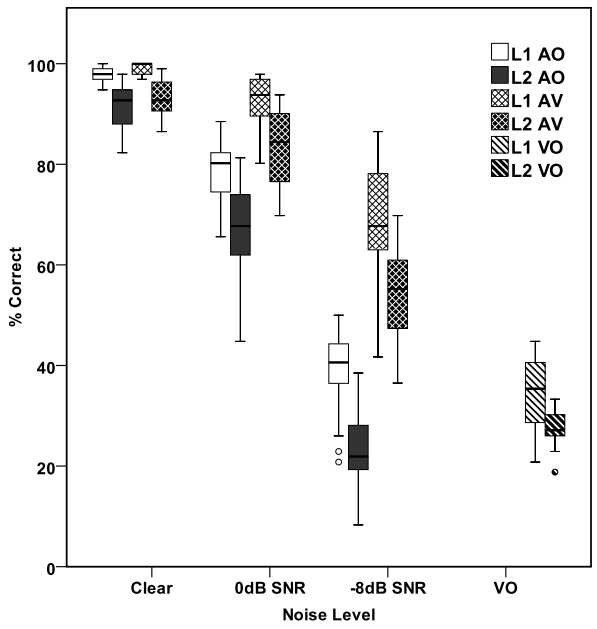

Figure 1. Percent correct scores for the auditory-only (AO), auditory-visual (AV) and visual-only (VO) conditions. Scores are depicted for both L1 and L2 listeners across across the three noise levels (clear, 0dB SNR, and -8dB SNR). Error bars are +1 standard error.

Auditory-Visual (AV) conditions

The percentage correct data for the AV conditions are also presented in Figure 1: L1 listeners consistently performed better than L2 listeners, and also the performance of both groups decreased as the noise levels in the stimuli increased. To assess the L1 and L2 differences, the AV percentage correct scores were analysed with a mixed repeated measures ANOVA as in the AO conditions.

The results showed that overall performance was greater for L1 than for L2 listeners ($F(1, 18) = 48.64$, $p < .001$, partial $\eta^2 = .73$) and that the performance of both groups decreased significantly as a function of increased noise level ($F(2, 36) = 380.77$, $p < .001$, partial $\eta^2 = .95$). The between-subjects main effect comparing the English and Spanish participants was again not significant ($F(1,18) = 2.89$, $p = .106$, partial $\eta^2 = .14$).

The primary interest in the AV scores was whether the inclusion of visual speech made the perception of L2 phonemes more robust in noise conditions. This would be supported by the absence of an interaction between Listener group and Noise level. However, similar to the AO percentage correct data, there was a significant Listener group by Noise condition interaction $F(2, 36) = 11.17$, $p < .001$, partial $\eta^2 = .38$. Again, pairwise comparisons (with Bonferroni adjustments) were conducted and this revealed that the Listener group by Noise level interaction was significant between the Clear and -8dB SNR conditions ($F(1\ 18) = 18.48$, $p < .001$, partial $\eta^2 = .51$). As can be seen for the AV results in Figure 1, the L1/L2 difference increased as the noise levels increased (increasing from 7% in the clear, to 10% in 0dB SNR, and then to 14% in -8dB SNR conditions) indicating that presence of visual speech did not compensate for L2 listeners' vulnerability to noise.

As can be seen in Figure 1 (comparing the AO and AV results), the amount of AV benefit (i.e. the relative improvement in performance from the AO to AV condition) increased significantly as the noise levels increased. This is consistent with previous research indicating that the benefit provided by visual speech increases as the SNR decreased [e.g., 32]. A repeated measures ANOVA for the AV benefit scores was conducted with Participant language (native English speakers; native Spanish speakers) as a between subjects factor, and Noise level (Clear, 0dB SNR, -8dB SNR) and Listener group (L1, L2) as within subject factors. As expected, the main effect of Noise level was significant, $F(2, 36) = 252.55$, $p < .001$, partial $\eta^2 = .93$, indicating the amount of AV benefit increased as noise levels increased. However, neither the main effect of Listener group $F(1, 18) = 1.70$, $p = .21$, partial $\eta^2 = .01$ or the interaction of Listener group by Noise level $F(2, 36) = .95$, $p = .396$, partial $\eta^2 = .05$ was significant, indicating that the visual effect was similar across the listener groups.

VO Conditions

Figure 1 also shows the percentage correct data for the visual only (VO) conditions. A repeated measures ANOVA was conducted (using the same within and between groups as with the AO and AV analyses). The main effect of Listener group was significant ($F(1, 18) = 15.22$, $p = .001$, partial $\eta^2 = .458$) with L1 listeners (mean 34%) performing significantly better than L2 listeners (mean 26%). This superiority for the perception of L1 over L2 consonants in visual only conditions seems to in part contribute to the L1/L2 differences in the AV conditions.

Collectively, the results of the percentage correct scores show that across the AO, VO and AV conditions, and across the clear and noise levels, L1 listeners performed significantly better than L2 listeners. L2 vulnerability to noise was found for the AV as well as the AO conditions. Like L1 listeners, L2 listeners received benefit from the provision of visual speech (especially in noise conditions as evidenced by increasing AV benefit from clear to the 0dB SNR and -8dB

SNR conditions). However, such benefit did not interact with the L2 listeners' vulnerability in noise.

SINFA analysis

For each of the AO, AV and VO conditions, confusion matrices were constructed and SINFA analyses were conducted. Reference to TI instead of raw consonant correct scores can be particularly beneficial in that TI takes into account the entire range of correct and incorrect responses that a participant made. Furthermore, TI scores takes into account response biases (e.g., randomly guessing across the items will give a result of zero) [33]. In this way, TI scores allow for a fuller picture of the speech perception of the participants, and allows for a finer analysis of the type of information that participants are able to recognise [33]. The aim of the following analyses therefore was to examine whether the pattern of TI (for the articulatory-acoustic features of voicing, manner and place of articulation) as a function of noise level varied for the L1 and L2 listeners.

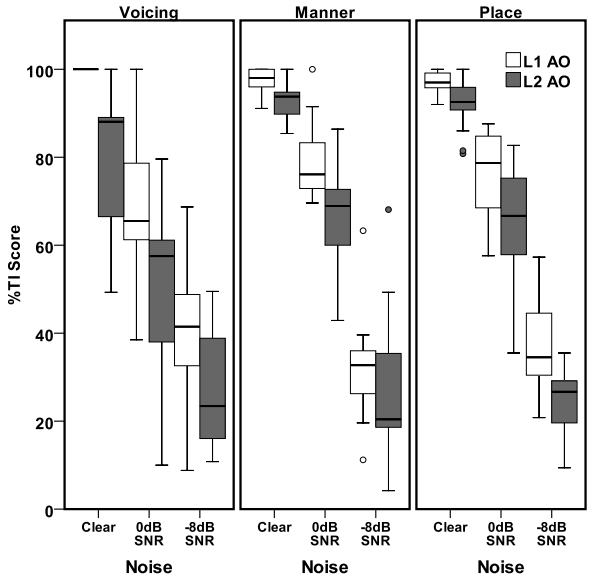

Figure 2. Transmitted information (TI) percentage scores for the auditory-only (AO) condition. TI percentages are shown for the L1 and L2 listeners across the three noise conditions (clear, 0dB SNR, -8dB SNR), for the features of voicing, manner and place of articulation. Error bars are +1 standard error.

Figure 2 displays the relative TI for the AO conditions. As can be seen, the broad patterns of performance established in the previous analyses are reflected in the TI scores for the L1 and L2 listener groups. That is, L1 listeners perceived consistently more information than L2 listeners and as the signal-to-noise ratio for the stimuli decreased, the amount of information transmitted decreased for both L1 and L2 listener groups. However, the largest L1 and L2 difference can be seen for the feature of voicing. Across the three noise levels, L1 listeners reflected the transmission of voicing information of 95.03% in clear, 67.07% in 0dB SNR, and 40.38% in -8dB SNR conditions. However transmission of voicing information for the L2 listeners was 75.94% in clear, 47.14% in 0dB SNR, and 24.67% in -8dB SNR conditions. This lower perception of voicing contrasts for L2 listeners is therefore at least in part driving the differences observed between L1 and L2 listeners for overall phoneme perception.

However, even accounting for the reduction in use of voicing information for L2 listeners, there is no particular property that shows a disproportionate reduction from clear to noise levels for L2 listeners: across all three features of voicing, place and manner, L1 and L2 listener's performance reduced at comparable rates from clear to the two noise conditions. Repeated measures ANOVAs with Listener group (L1; L2) and noise level (clear; 0dB SNR; -8dB SNR) as within-group variables were conducted for each feature individually. The main effects of Listener group was significant across each feature class ($p < .001$), as was the main effect of Noise level ($p < .001$) confirming the main effects of Listener group and Noise level evident from Figure 2. Interestingly, there were no significant Listener group by Noise level interactions for either voicing, manner or place; that is, the L1/L2 difference did not increase as a function of noise level for any individual feature assessed. As such, it appears that L2 vulnerability to noise indicated by the percentage correct scores was not due to the deterioration of particular features but to a general decline in the ability to select the correct segment in a feature class.

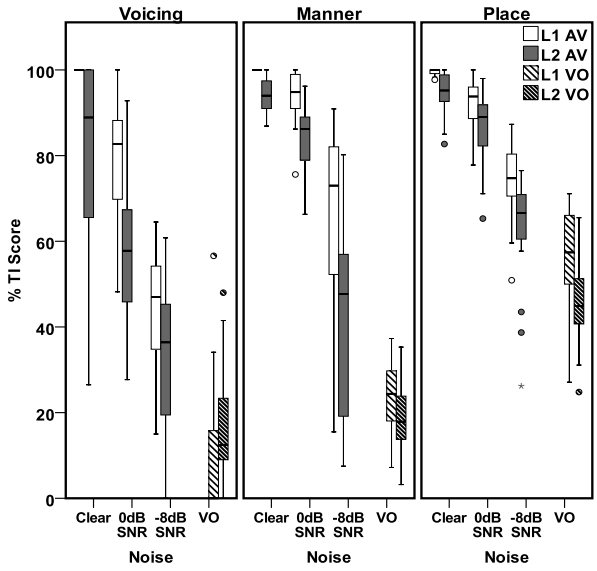

Figure 3. Transmitted information (TI) percentage scores for the auditory-visual (AV) and visual-only (VO) conditions.

Figure 3 displays the TI scores for the various AV conditions and for the VO condition. As with the percentage correct data, scores were consistently greater for L1 listeners than for L2 listeners, and the TI percentage decreased as a function of increasing noise level. Similarly to the AO conditions, the largest difference between L1 and L2 listeners across the three noise levels was clearly for voicing. This finding is not surprising given that visual speech primarily conveys information to place of articulation and relatively little information to voicing [e.g., 34], the provision of visual speech was unlikely to attenuate any L1/L2 discrepancy in the perception of voicing information. Indeed, the VO data showed that place information was most accurately perceived, followed by manner, and then by voicing information.

For the AV data, repeated measures ANOVAs with Listener group (L1; L2) and noise level (clear; 0dB SNR; -8dB SNR) as within-group variables were conducted for each feature individually. As expected, across each feature class of voicing, manner and place of articulation, the main effects of Listener group ($p < .001$ across each feature) and Noise level ($p < .001$ across each feature) were significant. L1 listeners performing significantly better than L2 listeners, and both groups significantly decreasing as the noise levels increased. While the Listener group by Noise level interaction was not significant for voicing and place of articulation, it was significant for manner of articulation ($F(2, 36) = 9.49$, $p < .001$,

partial $\eta^2 = .35$). That is, L2 listeners' TI was disproportionally lower than L1 listeners as a function of noise level (the mean L1/L2 difference for TI of manner of articulation in AV condition increasing from 5% in clear to 9% in 0dB SNR, to 23% at -8dB SNR). Follow-up comparisons revealed that only the significant difference between the two listener groups was found only in the transmission of manner in -8dB SNR ($F(1, 18) = 19.94$, $p < .001$, partial $\eta^2 = .53$).

By comparing Figure 2 and Figure 3, it can be seen that both L1 and L2 groups received substantial AV benefit in noise conditions for the features of manner and place of articulation (i.e. the increased performance for AV, compared to AO conditions, was greater for manner and place of articulation scores than it was for voicing). Again, this is consistent with prior literature, i.e., the greatest AV benefit can be expected for conditions where the auditory and visual cues provided are complimentary [e.g. 34]. As visual speech predominantly provides cues to place, and to a lesser extent, manner of articulation, the greatest AV benefit will be expected in those conditions, as opposed to voicing, where little benefit is added by providing visual speech cues [34]. As mentioned above, the largest AV benefit (i.e. the largest L1/L2 discrepancy) was found for the perception of manner of articulation in -8dB SNR condition, indicating L2 listeners' relative vulnerability to noise.

In summary, the primary point of difference between the L1 and L2 groups for TI across both the AO and AV conditions was for voicing information. In AO conditions, there were no significant Listener group by Noise level interactions for either voicing, manner or place of articulation, indicating that no single feature accounted for the L2 vulnerability in noise effects found for the percentage correct data. However, an L2 vulnerability in noise was found in the perception of manner of articulation for the AV conditions.

Integration Efficiency (IE)

The analyses above showed that AV perception was greater for L1 than for L2 listeners and this was partly due to the L2 perceivers being less able to extract information from visual speech. However, it is not clear whether it was also due to L2 listeners' being less efficient in integrating auditory and visual information [e.g., 25]. In order to examine the integration efficiency of the participants, a model (PROB) proposed by Blamey et al (1989) was used. PROB assumes that AV speech recognition errors occur when simultaneous errors in both auditory and visual perception occur [31, 34]. Through comparing the error rates for features of speech perception (i.e., voicing, manner and place of articulation) for auditory and for visual conditions, the model can be used to predict the "optimum" performance in AV conditions with the ratio between participants' actual and predicted scores providing a measure of integration efficiency. Although researchers have claimed that it is not an "optimum" prediction model per se (for example, Grant, Walden, and Seitz, (1998) found that PROB occasionally underestimated AV performances), and also that it does not allow for AV speech integration occurring early and at a prelexical stage of processing [e.g., 15], due to its simplicity of application and that it incorporates information transfer rates rather than overall recognition scores, PROB was used in the current study as measure of integration efficiency.

The integration efficiency scores for the AV conditions (estimated as a ratio for predicted scores from PROB with the observed TI scores, expressed as a percentage: predicted/observed * 100) across the separate information transmission features (of voicing, manner and place of articulation) are detailed in Figure 4. Examining the predicted scores, it was evident that PROB often underestimated the participants' actual AV performances (i.e., as evidenced by integration efficiency scores above 100%). However, as the underestimated scores likely represent greater integration efficiency [34] they were used to compare the performance between the groups.

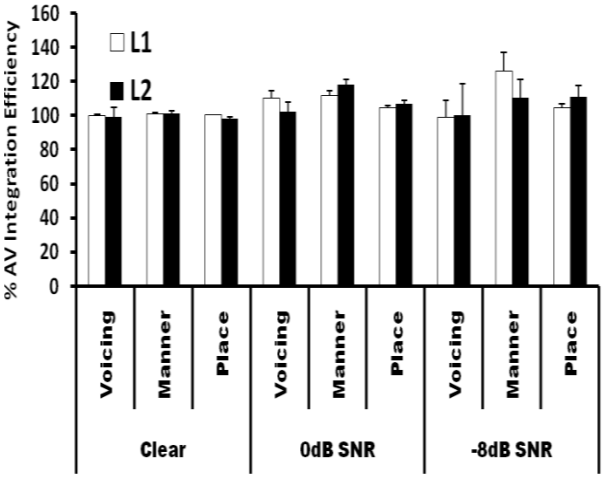

Figure 4. Auditory-visual (AV) Integration efficiency scores expressed as a ratio between the observed correct and the predicted correct recognition performances. IE scores are presented for both L1 and L2 listeners, across the three noise conditions (clear, 0dB SNR, -8dB SNR), for the features of voicing, manner and place of articulation. Error bars are +1 standard error.

A repeated-measures ANOVA was carried out with Participant language as a between subjects factor and Listener group, Noise level and TI (voicing, manner, place) as within-subjects factors. The main effect for Listener group was non-significant: ($F(1, 18) = .185$, $p < .672$, partial $\eta^2 = .01$). In contrast to previous research [e.g. 25], this result indicates that the L1/L2 differences in the ability to perceive AV speech information (detailed in the sections above) were not due to differences in their ability to integrate the available auditory and visual information. That is, once differences in the L1 and L2 listeners' ability to perceive speech information were accounted for, neither L1 nor L2 listeners were more or less efficient at integrating the available information in AV speech conditions. Interestingly this finding held across noise levels indicating that the L2 vulnerability to noise in AV conditions appears due to differences in extraction of speech information, and not to differences in integration efficiency.

CONCLUSION

The current study sought to examine L1 and L2 differences in speech perception in noise for AO and for AV conditions. Consistent with previous findings, for the AO presentation conditions, listeners were significantly worse at identifying L2 consonants in noise, than they were L1 consonants [e.g., 7, 8]. Further, although the identification of both L1 and L2 consonants improved significantly in AV compared to AO conditions (i.e. AV benefit), the L2 speech preception vulnerability to noise effect still remained for the AV presentation conditions. That is, the source of the L2 speech perception in noise vulnerability did not interact with the inclusion of visual peech.

The SINFA analyses revealed that the differences between L1 and L2 listeners in auditory and AV speech were mainly due to a lower perception of voicing information for L2 listeners. As visual speech is unlikely to provide strong cues for

voicing information (but does so for manner and place of articulation) [e.g., 34], it is unsurprising that provision of visual speech did not compensate for L2 listeners' vulnerability to noise. Furthermore, for AO speech perception, the disproportionate reduction in performance for L2 listeners in noise appeared to reflect a problem in selecting the correct segment within a feature class, since when scored at the feature level there was no L1/L2 difference as a function of noise level. For the AV presentation conditions, however, there was a difference in the ability of L1 and L2 groups to correctly perceive manner of articulation as a function of noise level for AV presentations. This suggests that the use of AV manner information may require extensive exposure to the language.

ACKNOWLEDGEMENTS

The second author acknowledges support from the Australian. Research Council (DP0666857 & TS0669874).

REFERENCES

1 Werker, J. F., & Lalonde, C. E. (1988). Cross-language speech perception: Initial capabilities and developmental change. Developmental Psychology , 24, 672-683.
2 Polka, L., & Bohn, O.-S. (1996). A cross-language comparison of vowel perception in English-learning and erman-learning infants. Journal of Acoustical Society of America (100), 577-592.
3 Burnham, D. (1986). Developmental loss of speech perception: Exposure to and experience with a first language. Applied Psycholinguistics (7), 206-240.
4 Kuhl, P. K., Conboy, B. T., Padden, D., Nelson, T., & Pruitt, J. (2005). Early speech perception and later language development: Implications for the "critical period.". Language Learning and Development (1), 237-264.
5 Florentine, M., Buus, S., Scharf, B., & Canevet, G. (1984). Speech reception thresholds in noise for native and non-native listeners. Journal of the Acoustical Society of America (75), s84.
6 Hazan, V., & Simpson, A. (2000). The effect of cue-enhancement on consonant intelligibility in noise: Speaker and listener effects. Language and Speech (43), 273–294.
7 Garcia Lecumberri, M. L., & Cooke, M. (2006). Effect of masker type on native and non-native consonant perception in noise. Journal of Acoustical Society of America , 4 (119), 2445–2454.
8 Cutler, A., Garcia Lecumberri, M. L., & Cooke, M. (2008). Consonant identification in noise by native and non-native listeners: Effects of local context. Journal of Acoustical Society of America , 124 (2), 1264-1268.
9 Mayo, L. H., Florentine, M., & Buus, S. (1997). Age of second-language acquisition and perception of speech in noise. Journal of Speech, Language and Hearing Research (40), 686-693.
10 Best, C. T. (1995). A direct realist view of cross-language speech, perception, in W. Strange (Ed.) Speech Perception and Linguistic Experience, pp. 171–204. Timonium, MD: York
11 Flege, J. E. (1995). Second language speech learning: Theory, findings and problems, in W. Strange (Ed.) Speech Perception and Linguistic Experience, pp. 233–277. Timonium, MD: York
12 McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264, 746-748.
13 Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd, & R. Campbell (Eds.), Hearing by Eye: The Psychology of Lip-Reading (pp. 3-51). London, UK: LEA.
14 Burnham, D. (1998). Language specificity in the development of auditory-visual speech perception. In R. Campbell, & B. Dodd (Eds.), Hearing by Eye II: Advances in the psychology of speechreading and auditory-visual speech (pp. 27-60). London: Erlbaum.
15 Green, K. P. (1998). The use of auditory and visual information during phonetic processing: Implications for theories of speech perception. In R. Campbell, & B. Dodd (Eds.), Hearing by Eye II: Advances in the psychology of speechreading and auditory-visual speech (pp. 3-25). London: Erlbaum.
16 Rosenblum, L. (2005). The primacy of multimodal speech perception. In D. Pisoni, & R. Remez (Eds.), Handbook of Speech Perception (pp. 51-78). Malden, MA: Blackwell.
17 Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. The Journal of the Acoustical Society of America, 26, 212-215.
18 Grant, K. W., & Seitz, P. F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. Journal of the Acoustical Society of America, 104, 2438-2450.
19 Massaro, D. W. (1998) Perceiving talking faces. Cambridge, MA: MIT Press; 1998.
20 Grant, K. W., Tufts, J. B., & Greenberg, S. (2007). Integration efficiency for speech perception within and across sensory modalities by normal-hearing and hearing impaired individuals. The Journal of the Acoustical Society of America, 121, 1164-1176.
21 Ortega-Llebaria, M., Faulkner, A., & Hazan, V. (2001). Auditory-visual L2 speech perception: Effects of visual cues and acoustic-phonetic context for Spanish learners of English. In D. W. Massaro, J. Light, & K. Geraci (Eds.), Proceedings of the International Conference on Auditory-Visual Speech Processing (ASVP '01) (pp. 149-154). Santa Cruz, CA: Perceptual Science Laboratory.
22 Wang, Y., Behne, D. M., & Jiang, H. (2008). Linguistic experience and audio-visual perception of non-native fricatives. Journal of the Acoustical Society of America, 124, 1716-1726.
23 Hazan, V., Sennema, A., Faulkner, A., & Ortega-Llebaria, M. (2006). The use of visual cues in the perception of non-native consonant contrasts. Journal of the Acoustical Society of America (119), 1740-1751.
24 Sekiyama, K., & Tohkura, Y. (1993). Inter language differences in the influence of visual cues in speech perception. Journal of Phonetics, 21, 427-444.
25 Sekiyama, K., & Burnham, D. (2008). Impact of language on development of auditory-visual speech perception. Developmental Science, 11, 306-320.
26 Massaro, D. W., Cohen, M. M., Gesi, A., Heredia, R., & Tsuzaki, M. (1993). Bimodal speech perception: An examination across languages. Journal of phonetics (21), 445-478.
27 Hazan, V., Kim, J., & Chen, Y. (submitted). Audiovisual perception in adverse conditions: Language, speaker and listener effects.
28 Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. Behavior Research Methods, Instruments, & Computers, 35, 116-124.
29 Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. The Journal of the Acoustical Society of America, 27, 98-116.
30 Wang, M. (1976). SINFA: Multivariate uncertainty analysis for confusion matrices. Behavior Research Methods & Instrumentation, 8, 471-472.

31 Blamey, P. J., Cowan, R. S. C., Alcantara, J. I., Whitford, L. A., & Clark, G. M. (1989). Speech perception using combinations of auditory, visual, and tactile information. Journal of Rehabilitation Research, 26, 15-24.
32 Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cerebral Cortex, 17, 1147-1153.
33 Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. The Journal of the Acoustical Society of America, 116, 3668-3678.
34 Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. The Journal of the Acoustical Society of America, 103, 2677-2690.