Auditory, visual, and auditory-visual spoken emotion recognition in young and old adults

Simone Simonetti[1], *Jeesun Kim*[2], *Chris Davis*[3]

[1 2 3] The MARCS Institute, University of Western Sydney

s.simonetti@uws.edu.au, j.kim@uws.edu.au, chris.davis@uws.edu.au

ABSTRACT

The study examined the recognition of emotional speech as a function of the clarity of expression, the modality of presentation, and participants' age (M_{age} = 19.8 vs. 73.9). Based on the results of a previous study, expression clarity was varied by selecting Auditory-Visual (AV) recordings of one actor who had well recognised expressions of anger, happiness, sadness, surprise, disgust, and neutral and one actor who did not. The young (n = 24) and older (n = 19) participants were presented these stimuli in Auditory-Only (AO), Visual-Only (VO), or AV format and made a forced-choice judgement on each. Older adults performed worse than younger ones for all presentation modalities except clear VO expressions. Importantly, whereas younger adults showed an AV benefit (AV > VO), older adults did not (showing a presentation mode by clarity interaction). The importance of varying signal clarity when investigating age effects was discussed.

1. INTRODUCTION

Emotion recognition is important for successful interpersonal communication [22]. Generally, research shows that younger adults' emotion recognition is more accurate than older adults. Considerable research has examined the impact of age on facial emotion recognition (using static images). However, for the last decade or so, a growing number of studies have examined older adults' recognition of vocal emotion expression, particularly in speech (i.e., emotion prosody). These studies have typically presented participants with semantically-neutral sentences or words produced with different expressions (e.g., anger, fear, sadness, disgust, surprise, and happiness) and found that older adults performed worse than younger ones for all vocal expressions (see [8] for a review).

Various explanations for this difference have been advanced and subsequently tested. For example, one proposal has been that the emotion recognition problems that the elderly experience may in part be due to the emotional material having lexical content. Here the argument is that such content triggers linguistic processes [12] which in turn consumes cognitive resources and interferes with emotion recognition processes [3]. To test this idea, studies have used pseudowords or filtered speech with the aim of determining whether under such conditions performance improves. For example, Demenescu and colleagues [7] presented older adults and younger adults with pseudowords in an attempt to reduce cognitive load. However, they still found that older adults were worse than younger ones when recognising disgust and anger. Similarly, Kiss and Ennis [17] found that older adults performed worse than younger adults when presented with happy, angry, sad, fearful, and neutral pseudowords. Moreover, they found that when older adults were presented with emotional words and pseudowords, they had similar recognition levels. In sum, it appears that even if lexical processing causes some interference with emotion prosody processing, this is not the main cause for older adults' problems (a conclusion consistent with findings that older adults still show worse emotion recognition than younger ones with low-pass filtered speech) [19, 20].

Even though the semantic processing hypothesis has little support, it points the way to a potentially useful research strategy, where to understand the nature of the problem one investigates conditions where the performance of older adults might improve. In this regard, the current study examined two factors: (1) supplementing the auditory with visual emotion signal; (2) examining different levels of the clarity of the emotion signal.

With respect to (1), it is expected that visual emotion information will enhance auditory emotion processing. It has been shown that linguistic prosody can be expressed visually [4, 6, 15]; and that this facilitates auditory prosody processing [5]. Indeed, for younger adults, auditory-visual (AV) emotion perception is better than auditory only (AO) or visual only (VO) perception [16]. Given this, the problems that older adults have identifying auditory expressions may be reduced with AV emotional expressions.

Few studies have investigated the impact of aging on the recognition of AV emotional expressions and the results have been inconsistent. For example, Hunter and colleagues [14] presented their participants with AV emotional expressions and found that younger and older adults had similar levels of recognition accuracy across all emotions (i.e., happy, angry, sad, fearful, disgusted, and surprised). In contrast, Lambrecht and colleagues [18], who also presented their participants with AV expressions, found that younger adults outperformed older adults for the expressions happy,

alluring, and anger (older and younger adults showed similar performance for disgust).

The inconsistency in the above results may be partially due to methodological differences involving stimulus selection. It is this possibility that motivated the selection of the second variable (2) of the current study. That is, some studies may have selected stimuli that conveyed clear emotional signals so that participants can easily discern the presented emotion. This can be problematic as the use of very clear, unambiguous emotional expressions can produce ceiling effects and hence no age group and/or emotion type differences. Those studies that have shown age differences may have used stimuli where the expressions were less clear and thus may have been more sensitive to potential differences between age groups. That is, less clear stimuli may result in more response confusions and such response patterns can indicate the emotions that participants most often confuse. For example, it has been reported that older adults frequently label vocal emotions as more positive than they actually are [9]. These and other differences between older and younger adults may be more pronounced when participants are presented with less clear rather than clear-cut expressions.

For the current study we introduced a speaker clarity variable with two actors: one who had clearly recognisable emotional expressions (clear speaker); the other whose expressions were harder to recognise (unclear speaker). We also used AV (as well as VO and AO) expressions of emotion. Overall, it was expected that the younger group would outperform the older group across all conditions. Further, it was expected that both age groups would perform better in the AV compared to the VO and AO conditions.

2. METHOD

2.1. Participants

Twenty-four younger (8 males, M_{age} = 19.8, range = 17-27) and 19 older (11 males, M_{age} = 73.9, range = 62-89) adults were recruited. Younger adults were students and received course credit; older adults received monetary reimbursement. All reported English as their first language except 3 older adults who learnt it as a younger age (approx. 4 years).

2.2. Stimuli

The stimuli consisted of video recordings of 2 male native Australian English speakers, uttering 8 Semantically Unpredictable Sentences [2]. The speakers portrayed facial and vocal expressions of anger, sadness, disgust, surprise, happiness, or neutral as they spoke each sentence. Recordings were manipulated to produce VO, AO, and AV stimuli.

The two speakers were selected out of the five speakers in Kim and Davis [16] based on the results of their emotion recognition task. One speaker was ‘clear’ and the other ‘unclear’ at expressing emotion, i.e., participants were more/less able to identify the speaker’s emotions than any others’. A female speaker (taken from [16]) was used to present 12 practice trials. Finally, at the commencement of each new or different speaker, two neutral expressions of that speaker were presented to act as a speaker specific calibration.

2.3. Procedure

Participants first completed a questionnaire detailing age, gender, languages spoken, etc. Participants were told they would see (VO), hear (AO), or see and hear (AV) a person talking first neutrally, then in various emotional expressions. They were asked to select (using the mouse) one of five options presented on the screen (angry, sad, disgust, surprise, or happy) they thought best described the previously presented emotional expression. They were told to use the speakers’ neutral expressions as a baseline to compare with subsequently presented emotions.

Participants were tested individually in a quiet room and were presented with three blocks (VO vs. AO vs. AV) of 80 trials each that consisted of 8 sentences, 5 emotions (angry vs. sad vs. disgust vs. surprise vs. happy), and 2 speakers (clear vs. unclear), resulting in a total of 240 trials. The order of the three blocks was counterbalanced across participants so that participants could receive the AO, VO, or AV block first. Further, two versions of the experimental list were created so that half of the participants received the clear speaker trials first for all three blocks. The presentation order of stimuli within each speaker block was randomised using the DMDX display and response collection software [11].

In each block for each speaker, participants were first presented with two neutral expressions of the speaker, followed by the 40 items. For each trial, a fixation point was presented for 50ms, followed by an experimental item (approx. 6 secs). Then, 5 boxes labelled “Angry”, “Sad”, “Disgust”, “Surprise”, and “Happy” were presented on the screen for a response.

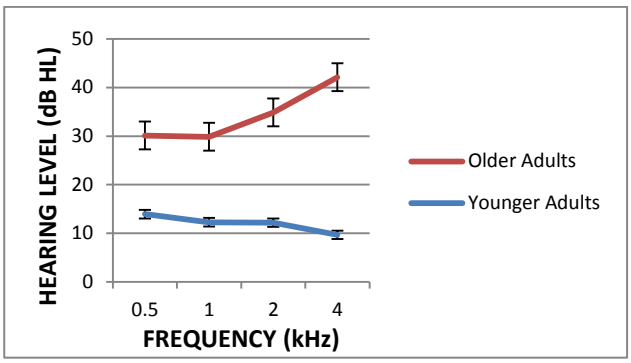
At the conclusion of the experiment participants completed a visual acuity test [1] and the Mini-Mental State Exam (MMSE) [10]. The MMSE was administered to check for signs of dementia as the presence of dementia is strongly associated with poor emotion processing and may be a confounding variable [21]. Hearing level was also assessed using pure tone audiometry (Diagnostic Audiometer, AD229e) for 4 different frequencies (0.5, 1, 2, and 4 kHz). Hearing ability can decline with age [23] and as such may interfere with the processing of acoustic signals that are important for emotion identification. Participants were then debriefed regarding the purpose of the study.

3. RESULTS

3.1. Visual Acuity, MMSE, and Hearing Ability

Hearing ability was averaged across both ears for each participant and across all participants for each frequency (Figure 1). Across all (particularly higher) frequencies, young performed better than older adults.

Figure 1: Hearing ability (dB) for each age group averaged across participants for 0.5, 1, 2, and 4 kHz. Error bars indicate standard error.



The results of the MMSE revealed that younger (M_{MMSE} = 29.08) and older (M_{MMSE} = 28.32) groups scored within the normal range (above 23) indicating no presence of dementia. Most participants had normal vision; one older adult showed slightly worse than normal vision with a visual acuity of 0.96 (where a score of 1 represents normal acuity).

3.2. Age Differences in Percent Correct Responses

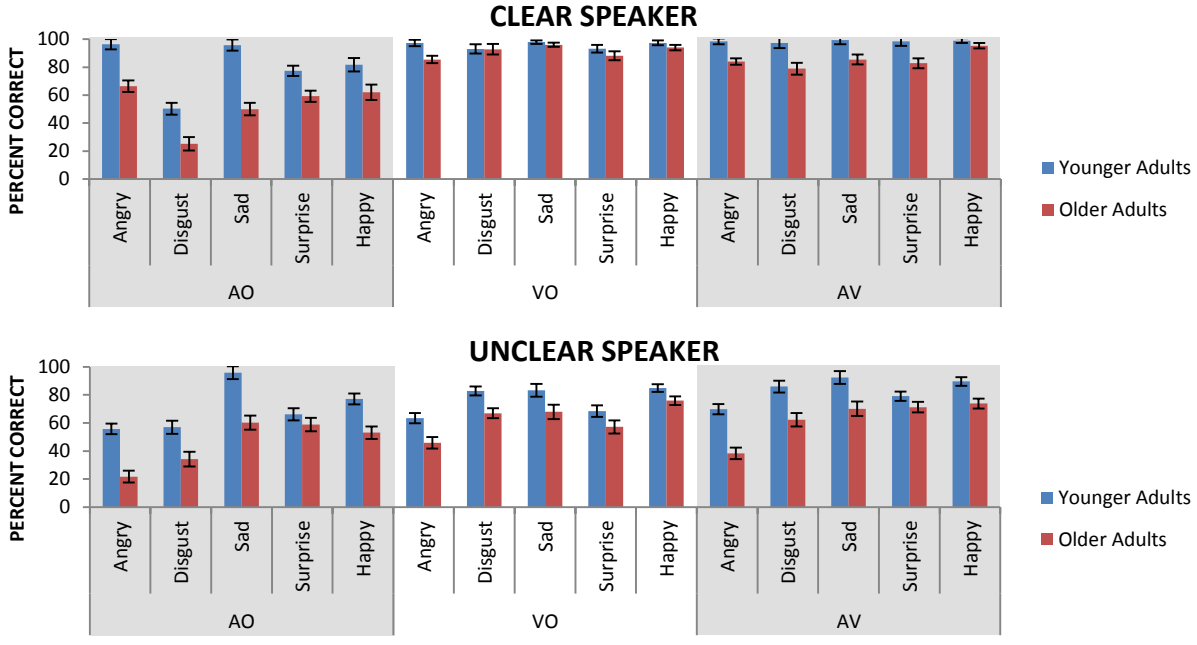
Percent correct responses were analysed in a mixed repeated measures ANOVA with presentation modality, speaker clarity, and emotion type as within subjects factors and age as the between subjects factor. Overall, the younger group (84.2%) outperformed the older group (66.1%) ($F_{(1,41)}$ = 48.41, p < .001, η_p^2 = .54). The VO (81.6%) and AV (82.6%) conditions attracted higher percent correct scores than the AO (62.2%) condition ($F_{(2,82)}$ = 194.57, p < .001, η_p^2 = .83). Emotion was better recognised for the clear (84.0%) than the unclear speaker (67.0%) ($F_{(1,41)}$ = 414.55, p < .001, η_p^2 = .91). The effect of emotion type was also significant ($F_{(4,164)}$ = 25.34, p < .001, η_p^2 = .38). There was a significant interaction between speaker and age ($F_{(1,41)}$ = 7.06, p < .05, η_p^2 = .15).

3.3. Speaker clarity

Separate ANOVA's were conducted for each of the speakers as above. Identification scores are presented in Figure 2. For these analyses a Bonferroni adjusted alpha of .017 was used [13]. For the clear speaker, younger adults (91.6%) outperformed older adults (76.4%) ($F_{(1,41)}$ = 35.57, p < .001, η_p^2 = .47). Performance was worse for the AO (66.5%) than the VO (93.6%) / AV (91.9%) conditions ($F_{(2,82)}$ = 270.14, p < .001, η_p^2 = .87). The effect of emotion type was significant ($F_{(4,164)}$ = 25.31, p < .001, η_p^2 = .38).

For the unclear speaker, younger adults (76.8%) performed better than older adults (57.2%) ($F_{(1,41)}$ = 51.80, p < .001, η_p^2 = .56). The VO (69.7%) and AV (73.3%) conditions attracted higher percent correct scores than the AO (58.0%) one ($F_{(2,82)}$ = 55.43, p < .001, η_p^2 = .58). The effect of emotion type was significant ($F_{(4,164)}$ = 37.66, p < .001, η_p^2 = .48).

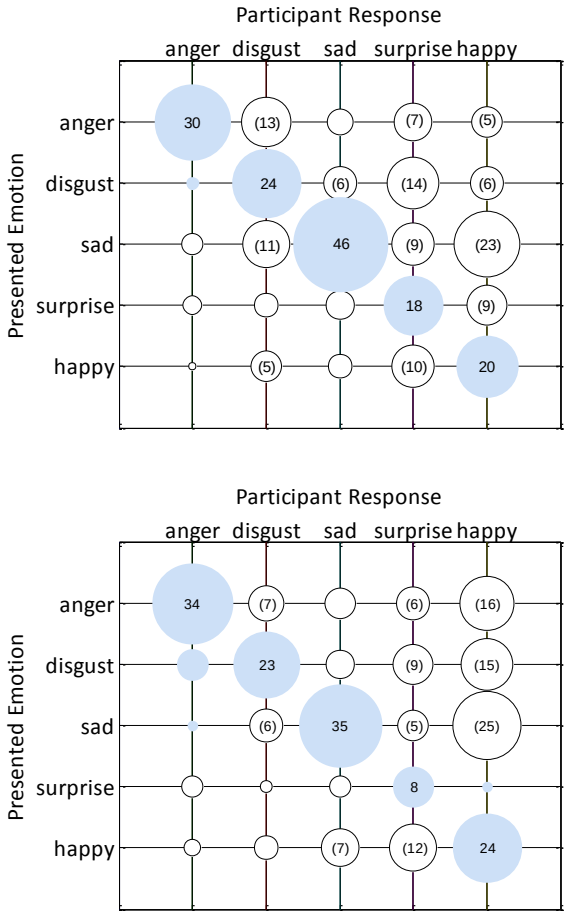
Figure 2: Mean percent correct identification scores for each emotion across AO (left panel), VO (middle panel), and AV (right panel) for the clear (top) and unclear (bottom) speakers. The error bars represent standard error.



3.4. Confusion matrices

We also looked at the age group difference in the AO condition by creating confusion matrices (Figure 3).

Figure 3: The difference in confusion matrices for the AO condition for the clear (top) and unclear (bottom) speakers. Older adult responses have been subtracted from younger adult responses. The numbers on the solid and transparent circles indicate the mean number of responses (%) for younger and older adults, respectively.



As can be seen, older adults had more confusions than younger adults. Older adults also showed different confusions for the clear and unclear speakers. Specifically, older adults showed more happy confusions (happy column) for the unclear (e.g., anger for happy, disgust for happy, sadness, for happy) than the clear speaker (e.g., anger for disgust, disgust for surprise, sadness for happy). This suggests that when older adults are presented with emotional expressions that they find ambiguous, they tend to perceive them as being more positive than they actually are [9].

4. DISCUSSION

The older adults of this study had elevated hearing thresholds compared to the younger ones and also performed much worse on AO emotion recognition. This is consistent with the proposal that hearing problems can affect auditory emotion recognition [18]. Two factors were examined that potentially could affect/enhance emotion recognition: (1) presenting AV emotional expressions and (2) the clarity of emotional signals. The results showed that overall, both groups performed better in the AV than the AO condition regardless of signal clarity. However, younger adults still outperformed older adults in the AV condition.

Interestingly, the clarity of the emotional signal appears to influence whether visual signals can be supplemented by auditory information. That is, in the clear condition, older adults performed better in the VO than the AV condition; whereas in the unclear condition older adults performed better in the AV than the VO condition. It is possible that when there is uncertainty in one modality with regards to the intended emotional expression (e.g., not-so-clear AO expressions) but not in the other modality (e.g., clear VO expressions), the older adults may become confused and thus less able to make use of the additional information presented in the AV condition. This particular finding may explain the inconsistent results reported between Hunter and colleagues [14] and Lambrecht and colleagues [18] where unclear auditory and clear visual emotional stimuli may result in age differences (and clear auditory and visual emotional stimuli may not).

In line with this, it is also possible that the signal clarity of VO expressions can influence age differences. Specifically, older adults showed worse performance than younger adults when presented with unclear VO emotional expressions but showed similar performance when presented with clear expressions. These findings show the importance of considering signal clarity when selecting research stimuli.

In conclusion, older adults experience problems when recognising auditory emotions. Some of these problems may be due to hearing loss [18]. The hope is that such problems could be offset by visual cues. Here, however we have shown that poor signal clarity can nullify any AV facilitation effect. It is therefore important that this factor is considered in future research.

5. REFERENCES

[1] Bach, M. 2006. The Freiburg visual acuity test – Variability unchanged by post-hoc re-analysis. *Graefe's Archive for Clinical and Experimental Opthamology* 245, 965–971.

[2] Benoît, C., Grice, N., Hazan, V. 1996. The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. *Speech Communication* 18, 381–392.

[3] Bowers, D., Coslett, H. B., Bauer, R. M., Speedie, L. J., Heilman, K. M. 1987. Comprehension of emotional prosody following unilateral hemispheric lesions: Processing defect versus distraction defect. *Neuropsychologia* 25, 317–328.

[4] Cvejic, E., Kim, J., Davis, C. 2010. Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion. *Speech Communication* 52, 555–564.

[5] Cvejic, E., Kim, J., Davis, C. 2011. Perceiving visual prosody from point-light displays in auditory-visual speech processing. In *Auditory-Visual Speech Processing 2011*.

[6] Cvejic, E., Kim, J., Davis, C. 2012. Recognising prosody across modalities, face areas and speakers: Examining perceivers' sensitivity to variable realisations of visual prosody. *Cognition* 122, 442–453.

[7] Demenescu, L. R., Mathiak, K. A., Mathiak, K. 2014. Age- and gender-related variations of emotion recognition in pseudowords and faces. *Experimental Aging Research* 40, 187–207.

[8] Dupuis, K. L. 2011. *Emotion in speech: Recognition by younger and older adults and effects of intelligibility (Doctoral dissertation)*. https://tspace.library.utoronto.ca/bitstream/1807/31738/1/Dupuis_Katherine_L_201111_PhD_thesis.pdf

[9] Fecteau, S., Armony, J. L., Joanette, T., Belin, P. 2005. Judgement of emotional nonlinguistic vocalisations: Age-related differences. *Applied Neuropsychology* 12, 40–48.

[10] Folstein, M. F., Folstein, S. E., McHugh, P. R. 1975. Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research* 12, 189–198.

[11] Forster, K. I., Forster, J. C. 2003. DMDX: A Windows display program with millisecond accuracy. *Behaviour Research Methods, Instruments, and Computers* 35, 116–124.

[12] Hanson, V. 1985. Cognitive processing in reading: Where deaf readers succeed and where they have difficulty. In Martin, D. (ed), *Cognition, Education and Deafness*. Washington, DC: Gallaudet University Press, 108–110.

[13] Hills, S. 2011. *Foolproof guide to statistics using IBS SPSS* (2nd ed.). Frenchs Forest, Sydney: Pearson Australia.

[14] Hunter, E. M., Phillips, L. H., MacPherson, S. E. 2010. Effects of age on cross-modal emotion perception. *Psychology and Aging* 25, 779–787.

[15] Kim, J., Cvejic, E., Davis, C. 2014. Tracking eyebrows and head gestures associated with spoken prosody. *Speech Communication* 57, 317–330.

[16] Kim, J., Davis, C. 2012. Perceiving emotion from a talker: How face and voice work together. *Visual Cognition* 20, 902–921.

[17] Kiss, I., Ennis, T. 2001. Age-related decline in perception of prosodic affect. *Applied Neuropsychology 8*, 251–254.

[18] Lambrecht, L., Kreifelts, B., Wildgruber, D. 2012. Age-related decrease in recognition of emotional facial and prosodic expressions. *Emotion* 12, 529–239.

[19] McCluskey, K. W., Albas, D. C. 1981. Perception of the emotional content of speech by Canadian and Mexican children, adolescents, and adults. *International Journal of Psychology* 16, 119–132.

[20] Mitchell, R. L., Kingston, R. A., Barbosa Bouças, S. L. 2011. The specificity of age-related decline in interpretation of emotion cues from prosody. *Psychology and Aging* 26, 406–414.

[21] Rosen, H. J., Wilson, M. R., Schauer, G. F., Allison, S., Gorno-Tempini, M., Pace-Savitsky, C., Kramer, J. H., Levenson, R. W., Weiner, M., Miller, B. L. 2006. Neuroanatomical correlates of impaired recognition of emotion in dementia. *Neuropsychologia* 44, 363–373.

[22] Surcinelli, P., Codispoti, M., Montebarocci, O., Rossi, N., Baldaro, B. 2006. Facial emotion recognition in trait anxiety. *Journal of Anxiety Disorders* 20, 110–117.

[23] Van Eyken, E., Van Camp, G., Van Laer, L. 2007. The complexity of age-related hearing impairment: Contributing environmental and genetic factors. *Audiology and Neurotology* 12, 345–358.