Cross-modality matching of linguistic prosody in older and younger adults

Simone Simonetti¹, Jeesun Kim², Chris Davis³

¹²³ The MARCS Institute, University of Western Sydney

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

Abstract

Older adults perform worse than younger adults in recognizing auditory linguistic prosody. Such problems may result from deterioration at the sensory level (e.g., hearing loss). The current study used a novel approach to examine this, by determining older adult’s performance on a visual prosody task. If older adults are able to process visual prosody this suggests that any difficulty they show when processing auditory linguistic expressions could be related to hearing loss. The current study presented 18 younger and 11 older adults with pairs of sentences spoken by the same talker. They decided whether the pair contained the same or different prosody in a simple AX matching task. Sentence pairs were presented in four different ways; auditory-auditory (AA), visual-visual (VV), audio-visual (AV), and visual-audio (VA). Compared to older adults, younger adults exhibited similar performance for focused statements but showed better performance for questions. We suggest that problems processing questioning expressions might result from hearing loss, problems perceiving pitch, or problems at the cognitive level.

1. Introduction

The linguistic meaning of an utterance can be determined not just by what is said (the content) but also how it is said (its prosody). To express prosody, talkers manipulate the acoustic features of an utterance, such as pitch, loudness, and duration. For example, talkers can raise pitch at the end of a sentence to indicate a question, and lower pitch for a statement [1]. Detecting changes in these acoustic properties is important for successful communication and difficulties in such processing can result in noticeable communication problems [2]. It is clear that simple hearing decline may result in problems when processing prosody. Given that older adults are known to have hearing decline and communication problems [3], it has been thought that they might have problems processing linguistic prosody information.

A number of studies have reported that compared to younger adults, older adults show poorer performance when labelling linguistic vocal expressions (e.g., questions, exclamations, commands, statements) [4, 5, 6]. It is possible that these reported differences result from age-related hearing loss. However, Mitchell and colleagues [5] reported that hearing sensitivity scores did not account for linguistic expression recognition scores. They suggested that older adults’ relatively poor performance may not just be due to sensorineural hearing loss but also may result from problems at other levels of processing (e.g., at the cognitive level).

The current study sought to better understand the processing level at which age-related prosody problems might occur. In this regard, the study tested older adults on their ability to process visual (as well as auditory) prosody in comparison to younger adults. Visual prosody refers to head and face movements that relate to and thus provide cues to auditory prosody [7]. For example, rigid head motion [8] and eyebrow movement [9] relate to changes in fundamental frequency whilst articulatory movements relate to changes in intensity [8]. Previous studies have demonstrated that people (younger adults) are sensitive to such cues [10, 11]. For example, [11] tested younger adults for their ability to match linguistic expressions (i.e., narrow focused statements, declarative statements, and echoic questions) within and across modalities. Participants were presented with two pairs of statements and were required to select the pair whose prosody matched (2 alternative forced choice, 2AFC, task). The results showed that younger adults were able to reliably match expressions within and across modality types, indicating their sensitivity to visual as well as auditory prosody cues. Further, the cross-modal matching results demonstrated that head and face movements are perceived as visual correlates of auditory prosody cues.

The rationale of comparing younger and older adults in the current study was that if older adults show no difficulty processing visual prosody then this would suggest that any difficulty they have when processing auditory linguistic expressions is related to hearing loss (i.e., at the sensory level); on the other hand, if older adults show difficulty processing visual prosody, it could indicate that they might be experiencing general prosody processing problems that go beyond the sensory level (e.g., cognitive processing). Note that although age-related decline in vision acuity is known, it is unlikely that this would cause problems for processing visual prosody. This is because a number of studies have reported no association between such decline and poor recognition of facial emotion [12, 13]. Indeed, our prediction was that like younger adults, older adults would have no problems in processing visual linguistic prosody. This prediction was based on our recent finding that older adults were able to recognise visual emotion prosody expressed while speaking just as well as younger adults, despite that they showed poor performance for the auditory counterpart [14].

To test the above, the current study adopted the method of [11] and examined the extent to which older and younger adults were able to match visual and auditory linguistic expressions (i.e., questions, focused statements) within and across modalities. Unlike [11], we used a simple AX matching task (instead of a 2AFC task) to minimise the possibility that older adults would show poor performance due to task demands or task complexity. For example, labelling tasks can reduce matching performance in older adults since such tasks might place a load on memory (which often declines in older age) [15]. For auditory-visual matching, we included auditory-visual (AV) and visual-auditory (VA) conditions. We did this because order of presentation has been shown to influence matching ability in younger adults interacting with relative cue strengths between channels [16]. Younger adults showed better performance in the AV than the VA modality possibly because auditory linguistic expressions provide more salient
cues than visual linguistic expressions [11, 16]. However, given that older adults might process visual linguistic prosody better than auditory linguistic prosody, there may be differences in AV and VA matching between the two age groups.

2. Method

2.1. Participants

2.1.1. Production participants

Two native female talkers of Australian English (Mage = 21.5) were recruited to record the audio and video stimuli. Both received monetary reimbursement.

2.1.2. Perception participants

18 undergraduate psychology students (3 males, Mage = 20.2, range = 17–32) from the University of Western Sydney and 13 older adults (7 males, Mage = 71.7, range = 61–81) participated in this study. Younger adults received course credit and older adults received monetary reimbursement. All were native speakers of English.

2.2. Stimuli

The stimuli consisted of audio and video recordings of twenty spoken sentences. These sentences were produced with linguistic prosodic expressions (see below) by two female talkers in their early twenties. A large number of IEEE sentence lists [17] were rated for emotional content on a Likert scale (using a 5-point scale from -2 = very negative, 0 = neutral, and +2 = very positive). The twenty sentences that received a score of 0 were selected.

The video and auditory recordings of the two talkers were captured individually in a sound attenuated booth over several sessions. Each talker was seated in front of a monitor that displayed each sentence one at a time. The video camera (Sony NCCAM HXR-NX30p) was situated directly above the monitor and captured video at 1920 x 1080 full HD resolution at 50 frames per second. The microphone (AT 4033a Transformerless Capacitor Studio Microphone) was placed approximately 20 cm away from the talkers’ mouth out of the cameras view. All audio recordings were sent through a Motu Ultralite mk3 audio interface with FireWire connection to a PC running CueMix FX digital mixer and then to Adobe Audition.

Talkers said aloud each sentence after first reading it silently. The talker was directed to look into the camera lens as they uttered the sentence. To produce these stimuli, talkers interacted with an interlocutor to elicit questions or narrow (corrective) focused statements. For a sentence with a narrow focus, talkers were required to correct an error produced by the interlocutor (Example 1).

Example 1
(a) Interlocutor: “He carved a BED\text{\textit{neutral}} from the round block of marble?”
(b) Talker: “He carved a HEAD\text{\textit{focused}} from the round block of marble.”

For an echoic question, talkers were required to question a narrowly focused statement produced by the interlocutor (Example 2).

Example 2
(a) Interlocutor: “Look in the corner to find the TAN\text{\textit{focused}} shirt.”
(b) Talker: “Look in the corner to find the TAN\text{\textit{question}} shirt?”

The recording session was blocked by the type of expression (neutral, focused, questioning) for the 20 sentences. Then the whole recording session was repeated on a subsequent day. Overall, each talker was recorded for a total of 120 sentences (20 sentences x 3 expression types x 2 repetitions).

Audio and video recordings were segmented into each sentence using MATLAB. Video recordings were stripped of audio and were cropped to include just the head area. Audio recordings were normalised so that the intensity of all sound files was equal to 60 decibels.

2.3. Experimental design

A same-different matching task was used to test perceivers’ sensitivity to different types of expressions in the auditory and visual modalities. Each trial presented a pair of the same sentence that were produced either with the same/congruent (focused-focused, question-question) or different/incongruent (focused-neutral, question-neutral) expression. It is important to note that the first item in each pair was always taken from the first repetition of the recording session; whereas the second item of each pair was always taken from the second repetition of the recording session. This was done to ensure that no tokens were repeated within the congruent trials.

The presentation modality of the pair varied in four different ways: auditory-auditory (AA); visual-visual (VV); auditory-visual (AV); and visual-auditory (VA).

There were 320 trials in total as each of the four presentation conditions included 80 trials that consisted of 2 talkers, 2 expressions (focused vs. question), 2 pair types (congruent vs. incongruent) and 10 sentences. Note that two sets of 10 different sentences were used for the AA and AV blocks and the VV and VA blocks to minimise any learning effects that may make the task easier. Given this, two versions of the experimental list were created so that half of the participants received half of the AA and AV sentences (sentences 1-10) and half of the VV and VA sentences (sentences 11-20). The other half of the participants received the opposite selection so that all 20 sentences were presented with each of the modality types.

2.4. Procedure

Participants were tested individually in a quiet room. Each was informed that she/he would be presented with a pair of sentences consisting of the same words in AA, VV, or AV form, but that half the time the way each sentence of the pair was expressed would be similar and half the time this would be different. If the participant judged the pair to be expressed in a similar way she/he should press the yes button otherwise, press the no button.

Participants were presented with four blocks (AA, VV, AV, VA) consisting of 80 trials each. The order of the four blocks was counterbalanced across participants so that participants could receive the AA, VV, AV, or VA block first.
The presentation order of stimuli within each block was randomised using the DMDX display and response collection software [18].

In each block, participants were first presented with 4 practice trials randomly selected from the experimental list (2 expressions x 2 pair types) followed by the experimental trials. For each trial, participants were presented with the first item, followed by the second item. Then, two options were presented on the screen (“no”, “yes”) for a keyboard response. Participants were given four breaks throughout the experiment.

Participants completed a visual acuity test [19] and a dementia screening test (Mini-Cog, [20]) as dementia is strongly associated with poor prosodic processing and may be a confounding variable [4]. 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 [21] and as such may interfere with the processing of acoustic signals that are important for identification. Participants were then debriefed as to the purpose of the study.

3. Results

3.1. Visual Acuity, Mini-Cog, and Hearing Ability

All participants had normal vision and scored within the normal range on the Mini-Cog test indicating no presence of dementia. Hearing ability was averaged across both ears for each participant and across all participants in each age group for each frequency (Figure 1). Across all (particularly higher) frequencies, young performed better than older adults.

![Hearing ability (dB) for each age group averaged across participants for 0.5, 1, 2, and 4 kHz. The error bars represent standard error.](attachment:image1.png)

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

3.2. Age Differences in Percent correct Responses

Percent correct responses were collapsed across sentences within each expression type. Participants with percent correct scores of less than 20% in any condition were removed. For the 18 younger adults and remaining 11 older adults (6 males, mean age = 72.4, range = 64–81), scores were collapsed across talkers within each expression type. These were analysed in a mixed repeated measures ANOVA with presentation modality (AA, VV, AV, VA) and expression type (focused, question) as within subjects factors and age as the between subjects factor. Identification scores are presented in Figure 2.

![Mean percent correct identification scores for each expression type across AA, VV, AV, and VA presentation modalities. The error bars represent standard error.](attachment:image2.png)

**Figure 1:** Mean percent correct identification scores for each expression type across AA, VV, AV, and VA presentation modalities. The error bars represent standard error.

Overall, there was no significant difference between the younger group (89.0%) and the older group (83.4%), $F(1,27) = 2.04, p = .17, \eta^2_p = .07$. The AA (89.3%), AV (86.6%), and VV (86.1%) conditions attracted the highest percent correct scores followed by the VA (81.0%) condition, $F(3,81) = 9.11, p < .001, \eta^2_p = .25$. Participants showed better performance for the focused statements (90.6%) than the questions (81.9%), $F(1,27) = 42.37, p < .001, \eta^2_p = .61$. The interaction between expression type and age group was significant, $F(1,27) = 14.61, p < .01, \eta^2_p = .35$. The interactions between modality and age group, $F(3,81) = 2.24, p = .09, \eta^2_p = .08$, and modality, expression type, and age group, $F(3,81) = .20, p = .90, \eta^2_p = .01$, were not significant.

The interaction between expression type and age group was analysed further using a simple effect comparison with a Bonferroni adjusted alpha. Simple comparisons reveal that younger and older adults showed similar performance when matching focused statements, $t(27) = .62, p = .89$, but that younger adults performed better than older ones when matching questions, $t(27) = 5.63, p < .05$. Furthermore, focused statements attracted higher percent correct scores than questions for both the younger, $t(27) = 4.76, p < .05$, and older, $t(27) = 4.29, p < .001$, adults.

4. Discussion

The main aim of the current study was to determine whether age-related linguistic processing problems would occur, and if so, at what level of processing. We found that older adults were able to match the prosody of focused statements but showed problems when matching questions, regardless of the presentation modality. Our interpretation of these findings is that because they could do well with one aspect of prosody (focus) older adults do not have a general problem processing linguistic prosody at the cognitive level. Older adults’ difficulties in processing prosody of questions (as well as other linguistic expressions such as statements, commands, and exclamations) plausibly results from specific problems at the sensory or perceptual levels, as discussed below.

In this study (as in other previous studies), older adults showed greater hearing loss than younger adults at higher frequencies than at lower frequencies. This age-related hearing loss may interfere with the detection of pitch change [22] such as the rise in pitch that is often contained at the end of a question [1]. Hence, it is possible that declines in pitch perception and hearing may be associated with poor
processing of auditory questioning expressions. This age-related hearing loss may also explain older adults’ poorer performance than younger adults’ when processing other linguistic expression types (i.e., statements, commands, exclamations) which typically contain changes in pitch but not in other prosodic cues such as word duration [1]. For focused statements which often contain duration changes at salient acoustic cues other than pitch [23], older adults performed as well as younger adults, indicating they had no problems in using duration cues.

The older adults’ relatively poor performance for visually presented questions may also be related to hearing loss. That is, poor hearing may result in a general bias to gaze longer at the mouth region than the eye region because the mouth can be used as a cue for speech perception [24]. Indeed, it has been reported that compared to younger adults, older adults tend to gaze more at the mouth than the eye region for emotional faces [25]. If so, older adults would miss useful visual cues of questioning expressions in the eye region. That is, pitch changes can be used to differentiate questions from statements [26] and the mouth region does not provide distinct cues related to pitch change [9]. Instead, visual changes associated with pitch change typically occur in the eye region (i.e., eyebrow raise, squinting of the eyes) [9]. On the other hand, duration cues which are useful for the detection of a focused statement are clear from the mouth region. These salient cues that are presented in the mouth region would be more likely to be detected by older adults who tend to gaze at the mouth region and therefore cues would be more accurately matched within and across modalities. Future research could track older and younger adult eye movements to determine whether older adults gaze more at the mouth when viewing linguistic expressions; or present these age groups with upper and lower face regions to determine whether older and younger adults show any differences in matching linguistic expressions in specific face regions.

An alternative explanation for the relatively poor performance of older adults’ can be given in terms of their limited cognitive resources and/or vulnerability to distraction. In this regard, it is important to note here that overall, all participants performed worse when matching questions than focused statements. This might be because questions lead to further processing whereas narrowly focused statements do not. That is, seeing or hearing a questioning expression possibly triggers a cognitive response in the listener as that is the natural function of a question [27]. This response then would consume cognitive resources and interfere with task performance (matching question prosody). Such interference would cost older adults more because they do not have as much cognitive resources as their younger counterparts [28]. Further, compared to young adults, older adults are reported to be more vulnerable to distraction and to have impaired sustained attention [29].

Note that this alternative explanation entails that older adults should be able to perceive questioning intonation. That is, to trigger a response to a question, the question first needs to be perceived (registered) and processed. For this to occur, processing at the sensory level needs to be intact. Future research could explicitly examine whether older adults are able to detect rises or falls in pitch similar to that used to signal questions. If they are able to do this, then this might indicate that the problem lies at the cognitive or perceptual level. If they are not able to do this, then this might suggest that the problem lies at the sensory and possibly also at the cognitive level.

Finally, older and younger adults performed better in the AV than the VA condition. Given that both groups performed better in AA than VV, this is consistent with the idea that order of presentation can impact matching performance interacting with the relative cue strength between channels. That is, matching performance is better when stable rather than unstable cues come first. It should be noted that the current analyses did not separate yes and no responses and further analyses of the current results are required. Hence, the conclusions discussed here are tentative and will be examined further.

5. References


