INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Perceiving Older Adults Producing Clear and Lombard Speech

Chris Davis[1], *Jeesun Kim*[1]

[1]The MARCS Institute, Western Sydney University, Australia

chris.davis, j.kim@westernsydney.edu.au

Abstract

We investigated the perceptual salience of clear and Lombard speech adaptations by older adults (OA) communicating to a younger partner in a diapix task. The aim was to determine whether these two speech styles are perceptually distinct (for auditory and visual speech). The communication setting involved either the younger partner only in babble noise (BAB_partner) or both talkers in babble noise (BAB_both). In the control condition (NORM), both talkers heard normally. To determine how perceptible OA adaptions to these noise conditions were, short (1-4 s) auditory only and visual only recordings of the OA talking to their partner were presented in two perception experiments. In Experiment 1, half of the OA stimuli were from the BAB_partner and half from the NORM condition; and participants were asked to judge whether the older adult was taking to a person who could hear them well or to someone who has trouble hearing them. In Experiment 2 participants decided between NORM and BAB_both stimuli. Participants did both sound-only and visual-only versions. Results showed both adaptations were perceived better than chance; the BAB_both condition was discriminated better from NORM than the BAB_partner one, and auditory judgements were better than visual ones (although these were correlated).

Index Terms: prosodic attitudes, speech prosody, expressive speech, auditory-visual speech

1. Introduction

When a talker communicates in circumstances that negatively influence speech intelligibility (e.g., when there is background noise; or when the partner has a hearing problem, etc), she/he will typically adopt a clearer speaking style (clear speech). The term 'clear speech' can be used to cover a wide range of speaking styles ranging from infant-directed to Lombard speech [1]. However, due to the differences in how Lombard is induced (by exposing the talker to noise, i.e., 'self-oriented'), usually this speaking style is distinguished from many of the others [2]. Moreover, it has recently been suggested that clear and Lombard speech involve quite distinct acoustic modifications [3].

The properties of clear speech are typically investigated by using speech spoken under 'normal' conditions as a contrast. So, for example, if clear speech was induced by speaking in a noisy environment, samples of this speech would be compared to those obtained in quiet conditions. Perceptual studies of clear speech typically investigate whether there is a clear speech intelligibility advantage, that is, whether speech produced in a clear speech style is more intelligible compared to normally produced speech.

In general, it has been found that clear speech is more intelligible than normal speech, although a clear speech advantage has not been consistently found for some listener groups and appears to be talker dependent [1]. Recently, there has been interest in whether older adults (OA) can produce a clear speech advantage. For example [4] took four short speech samples from 30 OA talkers produced in a problem-solving 'spot the difference' picture task (Diapix, [5]) either in normal, quiet conditions, or when communication partner had a simulated hearing loss (a situation that should induce 'interlocutor-oriented' clear speech). These samples were then intensity normalized, mixed with eight-talker babble noise (9 dB SNR) and presented in a listening experiment. It was found that OA could produce clear speech but that overall their speech was less intelligible than the normal and clear speech of younger adults (YA). Based on the finding about which acoustic properties were best associated with intelligibility scores, it was suggested that the poorer intelligibility of OA speech was due to less energy being present in a frequency range (1-3 kHz) that presents many important speech cues.

The current perception experiment examined clear, Lombard and normal speaking styles produced by older adults. However, rather than examine the intelligibility of clear and Lombard speech, we investigated whether people could determine whether a talker was producing these speech styles or not. That is, we simply asked participants whether they thought that the older adult talker was taking to a person who could hear them well or to someone who has trouble hearing them.

There are several reasons why it is interesting to examine the extent to which people can perceive whether a person is producing clear and/or Lombard speech or not. One reason is that human perceivers are excellent at combining and weighing different types of speech cue, so using perceiver judgments about the presence/absence of clear speech will possibly be a more sensitive index of the presence of these speech styles than measurements of specific acoustic properties. Moreover, such a task is firmly focused on the phenomenon of speech adaptation itself.

A second reason to be interested in the perception of whether or not a talker is producing these exaggerated speech styles concerns talker-listener interaction. One aspect of conversation under difficult circumstances that is seldom considered is the extent to which both partners perceive each other to be making an effort to facilitate communication. One may imagine that if the interlocutor is having difficulty perceiving a talker then she/he might expect that the talker would try to produce clearer speech. If it is perceived that an older talker is not making such an effort, then the interlocutor may feel dissatisfied and be less motivated to put effort into speech understanding. In this regard, the construct of 'listener

http://dx.doi.org/10.21437/Interspeech.2019-2210

motivation' that has recently been proposed in the cognitive-hearing literature [6] is apposite.

A final reason that motivates examining the perception of the presence/absence of clear and Lombard speech styles (rather than the intelligibility of these styles) is that with this judgment it is straightforward to use visual speech stimuli without having to deal with the issue of the large variation in participant's speech reading ability that occurs in visual speech intelligibility studies. Given that visual cues become more relevant in difficult communication settings, it is important to test how well the presence/absence of clear and Lombard speech styles can be perceived only from visual speech.

In the current study we used speech stimuli produced in a situation where interlocutor-oriented clear speech is typically produced, as well as stimuli produced in a situation in which Lombard speech is typical. For clear speech production, the communication partner was unable to clearly hear the talker (due to masking noise), but the talker themselves was not subject to this noise (like talking to someone with a mild hearing loss). In the Lombard situation, both interlocutors communicated in noise. We will examine judgments of the presence/absence of clear speech for both auditory-only and visual-only speech productions.

Under Lombard speech conditions speech production changes are in large part driven by the external environment (background noise); so it is likely that these changes will be uniformly produced by the OA talkers and perceivers will be able to distinguish speech produced in noise and in quiet for both the auditory-only and visual-only stimuli. The situation is less certain for the auditory only (AO) productions in the clear speech conditions. From [4], it would be expected that OA will produce clear speech in such conditions. However, since clear speech modification rely on the talker making an effort to provide the listener with more salient acoustic cues, these changes may be less uniform across talkers. Furthermore, [4] only examined the situation where the interlocutors could not see each other, whereas in our case we used speech tokens produced when the talker was looking at the listener (since we wanted to test visual speech as well as auditory speech). When visual speech information is available, it may be that OA talkers are less inclined to voluntarily make speech production modifications [7].

2. Method

2.1. Production study

2.1.1. Participants

From videos of 36 single sex pairs of native Southern English adult talkers between the ages of 19 and 84 years that participated in [8], videos of 12 pairs were selected. These pairs consisted of a primary older adult (OA) talker and a younger adult (YA) conversational partner (see below for selection criteria). The 12 OA had a mean age of 71.08 years (SD 5.2), consisted of 8 females and 4 males. Four had normal hearing (as defined by having a mean pure-tone hearing threshold <20 dB HL calculated over 0.25, 0.50, 1, 2 and 4 kHz) and 8 had mild age-related hearing loss (defined as having a mean threshold of 20-45 dB HL over .25-.50-1-2-4 kHz).

2.1.2. Recording setup

For the recordings, participant pairs were seated in adjacent sound-treated rooms, that were connected by a two-way window. Each participant wore an Eagle G157b lapel microphone and a pair of Vic Firth SIH-1 headphones. The primary talker (see below) was video-recorded using a 640*480 (VGA) camera at 24 fps.

2.1.3. Procedure

In the Diapix task, the OA was always assigned the role of a primary talker (Talker A) and the YA the secondary talker (Talker B). The primary talker (whose speech was analysed) was instructed to take the lead and do most of the talking. We used only young participants as Talker B based on findings that both young and older participants exhibit a higher level of social skill, as measured by the composite partner attention score, when paired with young adults than when paired with older adults [9].

Conversations between the pair of participants were elicited by having them engage in the spot the difference Diapix picture task. In this task, participants collaborate to find 12 differences without seeing their partner's picture. Participants were told to start in the top left-hand corner of the picture and work clockwise. The task was stopped after 10 minutes or after all differences had been found if earlier.

The Diapix task was conducted under three conditions relevant to this study: A no barrier (NORM) and two barrier conditions. In the NORM condition, both talkers heard each other normally. In one of the barrier conditions (BAB_partner), the secondary talker (Talker B) heard Talker A in 8-talker babble noise. The SNR for the BAB_partner condition was individually set using an adaptive procedure on the Modified Rhyme Test (MRT). In the other barrier condition (BAB-both), both talkers heard each other in 8 talker babble noise (at 0 dB SNR). Participants were always given the NORM condition first and the order of the two barrier listening conditions was randomized. Participants were given a different picture from the DiapixUK set [5] for each presentation condition. They were instructed that the pictures contained 12 differences and that they had a time-limit of 10 minutes to find these differences.

2.1.4. Data Processing

The auditory portion of the recordings were analysed using a set of customized Praat scripts. Mean intensity, fundamental frequency and mean energy in the 1-3 kHz band were measured as indices of vocal effort [10]. To obtain a measure of mean energy in 1-3 kHz (ME1-3kHz), pauses were removed from the speech samples of each talker and concatenated. The signal intensity was scaled to 75 dB and band-pass filtered (1-3 kHz) and the mean intensity of this band calculated.

2.2. Perception study

2.2.1. Participants

Thirty-two Year 1 Psychology students from the Western Sydney University were participants. 16 students (14 Female) participated in the clear speech experiment, BAB_partner vs. NORM (auditory-only and visual-only versions – see below) and 16 (13 Female) in the Lombard BAB_both vs. NORM experiment. Participants were all native speakers of English

(or had learned English from a young age) and had self-reported normal hearing and vision.

2.2.2. Item selection

The video recording (of Talker A) from each of the Diapix conditions was annotated using the ANVIL annotation editor [11] such that the time periods when the talker raised their head to look at their conversational partner were marked. Onset and offset times were then used to extract videos from each of the presentation conditions. Only videos that were between 0.75 s and 4.38 s were selected and there was no significant difference in length across the three production conditions [$F(2,256) = 1.88$, $p = 0.156$]. Also, we only selected videos that did not show overt hand gestures (as these could be distracting). Further, we selected videos where there were at least five tokens per person in each of the NORM, BAB_partner and BAB_both conditions. This resulted in the selection of a total of 136 recordings for the 12 different Talker A older adults for the NORM condition; 136 videos for the BAB_partner condition and 136 videos for the BAB_both condition.

2.2.3. Design

The perception study consisted of two separate experiments. 1) A clear speech perception experiment that consisted of BAB_partner versus Norm stimuli. 2) A Lombard perception experiment that consisted of BAB_both versus the Norm stimuli. For each of these experiments, there were auditory-only and visual-only versions.

2.2.4. Perception procedure

The clear speech (auditory-only and visual-only versions) and the Lombard speech (auditory-only and visual-only versions) experiments were run separately and participants were randomly assigned to one or the other. Participants were tested individually in a sound treated IAC booth. Participants did both the auditory-only and the visual only versions (the participation order was randomized).

Stimuli were presented using the DMDX software [12]. Stimuli from the 12 OA talkers were blocked by talker. Half of the trials in a talker block were from the NORM and half either from the clear speech condition or from the Lombard speech one. Participants were instructed that on each trial that they would be presented with a short sound or video (with no sound) clip of an older adult talking to someone. Participants were told that the task was to decide for each clip whether the older adult was talking to a person who can hear them well or to someone who has trouble hearing them. If the participant thought that the person was speaking someone who could hear them well, then were instructed to press a key labelled 'can hear OK'; if they though the person was talking as if the listener could not hear them well, they were to press a key marked 'cannot hear well'.

3. Results

3.1. Acoustic data analyses

3.1.1. Auditory-only NORM vs. BAB_partner stimuli

Measured over all stimuli (for all of the 12 talkers), there was a significant difference in mean energy between NORM (79. 5 dB) and BAB_partner (83.4 dB) stimuli, $t(1,11) = -4.839$, $p < .001$. Overall, no significant difference was found between mean energy in the 1-3 kHz band for NORM (67.1 dB) and BAB_partner (68.2 dB), $t(1,11) = -1.57$, $p = 0.145$. There was no significant difference in mean F0 between the NORM condition (202.9 Hz) and the BAB_partner condition (213.2 Hz), $t(1,11) = -0.86$, $p = 0.41$.

3.1.2. NORM vs. BAB_both stimuli

A significant difference was found in mean energy between NORM (79.5 dB) and BAB_partner (87.5 dB) stimuli, $t(1,11) = -9.14$, $p < .001$. A significant difference between mean energy in the 1-3 kHz band for NORM (67.1 dB) and BAB_partner (69.2 dB) stimuli was found, $t(1,11) = -3.19$, $p < 0.01$. Finally, a significant difference between mean F0 in the NORM (202.9 Hz) and the BAB-partner (242.1 Hz) stimuli was found, $t(1,11) = -3.22$, $p < 0.01$.

3.2. Perception data analyses

3.2.1. Auditory-only NORM vs. BAB_partner stimuli

Overall, participants made more errors in classifying the BAB_Partner auditory stimuli (42.24% errors) than they did in classifying the NORM stimuli (23.22% errors). These error data were analysed using a generalised linear mixed model (glmer package in R) with a binomial distribution (with participant and talker and random factors), and there was a significant difference between the two errors rates, z score = 2.52, $p < 0.01$. The high error rates for the BAB_partner stimuli indicate that participants misclassified these as NORM stimuli.

Overall, the errors for the BAB_partner auditory-only stimuli were lower than chance (50%), $t = -3.55$, $df = 191$, $p < 0.001$; as were the errors for the NORM stimuli, $t = -14.591$, $df = 191$, $p < 0.0001$. Even though overall the errors made on the BAB_partner stimuli were better than chance, there was considerable variation in the error rates of the stimuli produced by the different talkers. Indeed, only the error data from five of the 12 talkers, were significantly lower than chance (50%); errors from two talkers were significantly higher than chance and errors from the rest of the talkers did not differ from chance. Figure 1 shows this variability in the errors across each talker in the BAB_partner and NORM conditions.

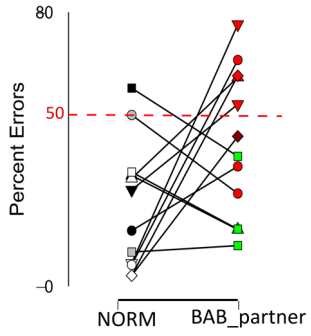

Figure 1: *Mean percent classification errors for the 12 Talker's Auditory-only NORM and BAB_partner stimuli. Chance performance is 50%.*

The BAB_partner stimuli from the five talkers who attracted the lowest error rates (16.7%, 16.6%, 35.0%, 27.1% and 11.9%) had greater mean energy than those from the four

talkers whose stimuli attracted the highest errors (66.1%, 61.5%, 61.0% and 76.4%), t(1,105) = 12.72, p < 0.001. Interestingly, three of the five talkers whose stimuli had the lowest error rates had normal hearing (green filled symbols), whereas all the talkers whose stimuli produced the highest error rates had mild hearing loss (red filled symbols).

3.2.2. *Visual-only NORM vs. BAB_partner stimuli*

The pattern of errors for the visual-only BAB-partner vs. NORM stimuli was similar to that for the auditory-only stimuli above. Overall, participants made more errors in classifying the BAB_Partner auditory stimuli (47.18% errors) than they did in classifying the NORM stimuli (31.84% errors). The data analysis using a generalised linear mixed model showed that the difference in errors was significant, z score = -2.271, p = 0.0232.

Unlike the auditory-only data, the BAB-partner error rate was not significantly different form chance (50%), t = -1.2031, df = 131, p = 0.2311. Figure 2 shows the error scores as a function of the items produced by each talker.

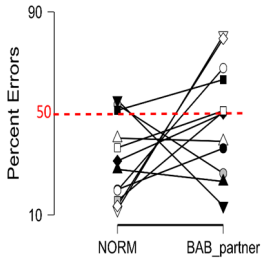

Figure 2: *Mean percent classification errors for the 12 Talker's Visual-only NORM and BAB_partner stimuli.*

The mean error rates for the visual speech stimuli (37.5%) from the five talkers whose auditory stimuli attracted better than chance performance (see above) was less than that for the visual speech stimuli from the four talker whose auditory stimuli had the worst error rates (59.4%).

3.2.3. *Auditory-only NORM vs. BAB_both stimuli*

. Figure 3 shows the error scores for the stimuli from each of the 12 talkers.

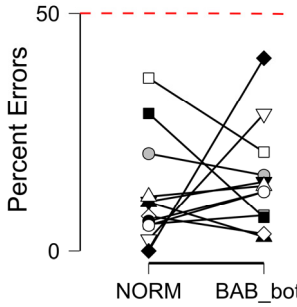

Figure 3: *Mean percent classification errors for the 12 Talker's Auditory-only NORM and BAB_partner stimuli*

There were far fewer errors in classifying the BAB_both (15.7%) and NORM (12.0%) stimuli. A generalised linear mixed model found that these error rates were not significantly different, z score = 0.98, p = 0.37; both error rates were lower than chance, t = -19.066, df = 107, p < 0.001; t = -26.169, df = 107, p < 0.001.

3.2.4. *Visual-only NORM vs. BAB_both stimuli*

Mean errors for the visual only BAB_both condition (37.7%) were higher than in the NORM condition (25.6%). A generalised linear mixed model found that these error rates were significantly different, z score = 1.96, p = 0.0498. This error pattern indicates that participants more often misclassified visual speech (that produced in noise) as being produced in quiet conditions. Both these error rates were lower than chance (50%), t = -6.0495, df = 107, p < 0.001; t = -13.765, df = 107, p <0.001.

Figure 4 shows the error scores for the stimuli from each of the 12 talkers. Stimuli for two of the talkers in the BAB_partner condition attracted error rates at chance level (50%) and for one talker at higher than chance levels (65.4%).

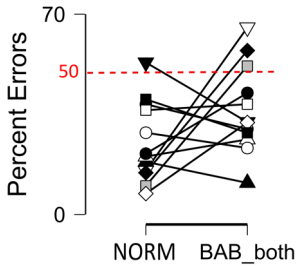

Figure 4: *Mean percent classification errors for the 12 Talker's Visual-only NORM and BAB_both stimuli.*

4. Discussion

The speech modifications that talkers make in various communication settings provide an insight into the dynamics of talker-listener interaction. We investigated the extent to which clear and Lombard speech could be perceived from OAs auditory and visual speech. We focused on the productions of older adults since our previous work suggests that their speech modifications are less effective than those made by younger adults (at least for situations when the talker themselves is not in noise) [8].

The findings of the current study indicate: 1). The perception of speech modification was better for the condition when both the talker and listener were in noise (Lombard speech) than when only the listener was (clear speech). This shows that the two speech styles are perceived differently and that older adults can produce salient speech modifications in response to the auditory environment but do so less to the needs of the interlocutor. 2). There was considerable variation in how the productions from the talkers were judged. The difference between talkers appeared to be related to the mean energy level used in the barrier conditions (especially in the BAB_partner condition). There was an intriguing hint that performance was poorer for the productions of those who had age-related hearing loss. 3). Overall, auditory productions were more salient for speech modification presence/absence judgments than visual ones. This could be a general property of these modalities or specific to OA productions; follow-up experiments with YA talkers will tell.

5. Acknowledgements

The authors acknowledge support by an ARC grant (DP150104600) and thank Outi Tuomainen and Valerie Hazan for collecting the video recordings and for comments on an earlier draft.

6. References

[1] R. M. Uchanski, D. B. Pisoni, and R. E. Remez, "Clear speech," in *The Handbook of Speech Perception* (Chapter 9. pp. 207-235), Blackwell Publishing, 2005.

[2] R. Smiljanić and A. R. Bradlow, "Speaking and hearing clearly: Talker and listener factors in speaking style changes," *Language and Linguistics Compass*, vol. 3, no. 1, pp. 236-264, 2009.

[3] E. Godoy, M. Koutsogiannaki, and Y. Stylianou, "Approaching speech intelligibility enhancement with inspiration from Lombard and clear speaking styles," *Computer Speech & Language*, vol. 28, no. 2, pp. 629-647, 2014.

[4] V. Hazan, O. Tuomainen, L. Tu, J. Kim, C. Davis, D. Brungart, and B. Sheffield, "How do aging and age-related hearing loss affect the ability to communicate effectively in challenging communicative conditions?" *Hearing Research*, vol. 369, pp. 33-41, 2018.

[5] R. Baker and V. Hazan, "DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs," *Behavioural Research Methods*, vol. 43, no. 3, pp. 761-770, 2011.

[6] M. K., Pichora-Fuller, S. E. Kramer, M. A. Eckert, B. Edwards, B. W. Hornsby, L. E., Humes, ... and G. Naylor, "Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL),". *Ear and Hearing*, vol. 37, pp. 5S-27S, 2016.

[7] M. Fitzpatrick, J. Kim, and C. Davis, "The effect of seeing the interlocutor on auditory and visual speech production in noise," *Speech Communication*, vol. 74, pp. 37-51, 2015.

[8] V. Hazan, O. Tuomainen, J. Kim, C. Davis, D. Brungart, and B. Sheffield, "Clear speech adaptations in spontaneous speech produced by young and older adults," *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1331-1346, 2018.

[9] D. D. Vandeputte, S. Kemper, M. L. Hummert, K. A. Kemtes, J. Shaner, and C. Segrin, "Social skills of older people: Conversations in same-and mixed-age dyads," *Discourse Processes*, vol. 27, no. 1, pp. 55-76, 1999.

[10] H. Traunmuller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children," *The Journal of the Acoustical Society of America*, vol. 107, pp. 3438–3451, 2000.

[11] M. Kipp, "Anvil-a generic annotation tool for multimodal dialogue," In *Seventh European Conference on Speech Communication and Technology*, 2001.

[12] K. I. Forster and J. C. Forster, "DMDX: A Windows display program with millisecond accuracy," *Behavioral Research Methods: Instruments & Computers*, vol. 35, pp. 116–124, 2003.