THE EFFECT OF VISUAL CUES ON SPEECH CHARACTERISTICS OF OLDER AND YOUNGER ADULTS IN AN INTERACTIVE TASK

Valerie Hazan [1], Outi Tuomainen [1], Jeesun Kim [2], Chris Davis [2]

[1] Department of Speech Hearing and Phonetic Sciences, UCL, UK; [2] MARCS Institute, Western Sydney University, Australia

v.hazan@ucl.ac.uk, o.tuomainen@ucl.ac.uk, J.Kim@westernsydney.edu.au, chris.davis@westernsydney.edu.au>;

ABSTRACT

This study investigated whether seeing a conversational partner while carrying out a collaborative task (diapix) in easy and difficult communicative conditions affected clear speech adaptations in older and young adults. 17 older (OA) and 13 young (YA) women were recorded while doing diapix with a conversational partner; in one condition, they could hear each other normally (NORM) while in another, the partner had a simulated hearing loss (HLS). Both conditions were in audio-alone and audiovisual modes. Articulation rate, fundamental frequency, mid-frequency energy and gaze count were analysed. In NORM, seeing their interlocutor had little effect on acoustic characteristics for OA or YA talkers. In HLS, across talker groups, gaze frequency increased and clear speech adaptations in articulation rate and mid-frequency energy, but not F0, reduced when talkers saw their partners. These findings support the view that interlocutor needs and the aim to minimise talker effort both affect clear speech adaptations.

Keywords: Speech production, speaker-listener interaction, clear speech, spontaneous speech, ageing

1. INTRODUCTION

Much communication occurs in challenging conditions. Communication difficulties can be due to environmental factors such as noise or the presence of other voices in the background. They can also be due to listener-related factors when, for example, communicating with someone who has a profound hearing loss (for a review see [12]). Much research in recent years has focused on the clear speaking style that speakers adopt in such conditions to maintain effective communication. In clear speech, adaptations are typically made to acoustic-phonetic features such as articulation rate, fundamental frequency mean and range, spectral tilt, vowel space (for a review, see [5]). A recent study has focused on the clear speech adaptations made by older adults and young adult controls when carrying out a problem-solving task with a conversational partner (a young adult) in good and challenging listening conditions [10]. Older adults (OA) had either normal hearing thresholds or mild age-related hearing loss. In challenging conditions, older adults with normal hearing typically patterned with young talkers while those with hearing loss made clear speech adaptations more consistent with an increase in vocal effort. This suggested that even a mild hearing loss in healthy OAs could affect clear speech adaptations.

These findings were for recordings made when participants could hear but not see each other. Face-to-face communication is beneficial in challenging conditions as it provides visual cues to conversational partners that complement auditory cues, as well as backchannelling cues (e.g., nodding) that can signal understanding. As suggested by Lindblom's Hyper-Hypo model of speech production [11], talkers will reduce speaker effort if this does not impact communication efficiency, so face-to-face communication (AV) could lead to a reduction in clear speech adaptations relative to when no visual cues are present (A) [8]. However, in a study comparing A to AV communication between participant pairs for a simple problem-solving task, adaptations to some suprasegmental aspects of speech were not significantly reduced in the face-to-face setting [9]. Here, this effect is explored with a less constrained task and a larger sample including older as well as younger talkers. In [6], an analysis of gaze frequency and duration suggested that older adults with normal hearing looked less and less often at their conversational partner, which could lead to an age effect in the A to AV comparison.

Our research questions were as follows:

- How does seeing an interlocutor in communicative speech affect the suprasegmental aspects of speech production in easy and difficult communicative conditions?
- What is the impact of speaker age on this effect of face-to-face communication?
- Is there a correlation between gaze frequency and degree of acoustic adaptations?

We hypothesised that (a) there would be a higher gaze frequency for older adults with hearing loss than for their hearing peers or young adults due to their own increased reliance of visual cues and (b) greater gaze frequency would be related to a decrease in acoustic adaptations.

2. METHOD

2.1. Participants

Participants were a random subset of the adult female talkers recorded for the elderLUCID corpus [10]. Thirty female talkers ('Talker A') of Southern British English were divided into two age groups: 'older adults' (OA) (*N*=17, range: 64-77 years, *M*=70.7) and 'younger adults' (YA) (*N*=13, range: 19-26 years, *M*=21.4). Participants reported no history of speech or language impairments and normal or corrected vision. YA participants all had normal hearing thresholds. Within the OA group, 9 OANH (*M*=69.3 years) had normal hearing defined as a mean pure-tone hearing threshold <20 dB HL calculated over .25-.50-1-2-4 kHz (mean better ear average: 11 dB HL, *S.D.* 6.1) while 8 OAHL (*M*: 72.2 years) had a mild age-related hearing loss defined as a mean threshold of 20-45 dB HL (mean better ear average: 28.4 dB HL, *S.D.* 5.4). The OA and YA groups did not differ significantly in background cognitive measures (digit span and word association tests).

A further 30 young women (range: 18-30 years; *M*=21.0) were conversational partners ('Talker B') whose speech was not analysed. Participant pairs did not know each other prior to testing.

2.2. Procedure

2.2.1. Experimental task

Diapix [14], a problem-solving 'spot the difference' picture task, was used (with the diapixUK picture pairs [2]) to elicit spontaneous interactions in a communicative situation between the pair of participants. Participants had to collaborate to find 12 differences without seeing their partner's picture. Talker A (whose speech was analysed) was told to lead the conversation. 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.

2.2.2. Recording conditions

Before the recordings, each participant pair practised diapix for 5-10 minutes while seated in the same room. For the recordings, they were seated in adjacent sound-treated rooms, connected by a two-way window. They wore Eagle G157b lapel microphones and Vic Firth SIH-1 headphones.

In addition to the condition where both talkers interacted without any interference (NORM), diapix was carried out when communication was impaired for one or both participants, to naturally elicit a clear speaking style. Here, data are presented for the hearing loss simulation (HLS) condition, where Talker B had a simulated severe-to-profound hearing loss (see [10] for full description). Participants were told that their partner had a simulated hearing loss but they did not experience this directly.

The NORM and HLS conditions were carried out in two modes: audio (A), where talkers could only hear each other, as a window blind was pulled down, and audiovisual (AV) where they could also see each other via a window. Recordings in A and AV modes were done at separate sessions, with the mode order randomised across participants. At each session, NORM was first, and the adverse conditions randomised within groups. Each talker was recorded on a separate channel (16 bit, 44,100 Hz sampling rate). In AV, Talker A was video-recorded with a 640*480 (VGA) camera at 30 fps.

2.3. Data processing

For all recordings, a cloud-based speech recognition system [4] was used to obtain time-aligned orthographic transcriptions of each channel. These were manually checked and corrected for orthographic and word-alignment errors. The following acoustic characteristics were analysed for Talker A's speech recordings. For further details about analysis procedures, see [10].

Articulation rate was calculated as the number of syllables produced by Talker A, calculated using the qdap package in R [13], divided by the total duration (in seconds) of the speech regions.

For fundamental frequency (F0) measures, for each file, a Praat [3] script was used to concatenate speech intervals and F0 calculations were done using the 'pitch' function. We used a formula [7] to calculate ceiling and floor limits specific to each talker, to exclude rogue values. For each talker, median F0 values and F0 range (1^{st} to 3^{rd} quartile) were calculated per condition and mode.

For the mean energy 1-3 kHz (ME1-3kHz) measure, long-term average spectrum analyses were done using a Praat script. After excluding speech segments above 88 dB, segments were concatenated and signal intensity scaled to 75 dB. The signal was band-pass filtered (1-3 kHz) and the mean intensity of this band calculated relative to the total energy. An increase in the relative energy in this mid-frequency band reflects a reduction in spectral tilt.

ANVIL AV annotation software [1] was used to calculate gaze frequency. An annotator marked when Talker A raised their head to look at their conversational partner. For each talker, gaze count and total gaze duration were calculated per condition.

3. RESULTS

Repeated-measures ANOVAs were carried out with a between-subject factor of group and within-subject factors of condition (NORM, HLS) and mode (A, AV). Initially, the group effect was investigated splitting OAs according to their hearing status. Where the OA groups did not differ significantly, statistics are presented comparing all OAs to the YA group.

3.1. Gaze frequency

Talkers looked at their interlocutor more often (Table 1) in HLS than NORM, $F(1,26)=121.3$, $p<.001$, $\eta p^2=.82$, but neither group ($p=.454$) nor group*mode interaction ($p=.142$) effects were significant, although there was a trend for YAs to have a higher gaze count in NORM. The large variance and significant correlation across conditions ($r(28)=.662$, $p<.001$) suggest that gaze frequency is a talker characteristic. The same effects were obtained for gaze duration. As count and duration were correlated (all $p<.001$), gaze count is used in further analyses.

Table 1: Mean and standard deviation for gaze count and duration (in seconds) in AV. There are missing data points for two OA talkers.

Group	NORM		HLS	
	Count	Dur.	Count	Dur.
YA (N=13)	35.6 (24.5)	48.3 (35.6)	79.2 (18.0)	153.8 (59.4)
OA (N=17)	21.2 (20.9)	24.3 (27.9)	80.0 (41.6)	126.4 (72.9)

3.2. Effect of visual cues on suprasegmental characteristics in NORM condition

Next, we investigated whether, when communicating in good listening conditions, seeing the conversational partner had a different effect on the speech produced by YA and OA talkers. We predicted that the absence of visual cues in A would lead OAs to produce clearer speech in this condition, thus leading to a smaller A to AV change in suprasegmental speech characteristics, as would be shown by a group*mode interaction.

For articulation rate (Fig. 1), there was no significant effect of mode ($p=.260$) or group*mode interaction ($p=.321$). YAs ($M=3.93$ syll/s) spoke faster than OAs ($M=3.56$), $F(1,27)=13.22$, $p=.001$, $\eta p^2=.33$.

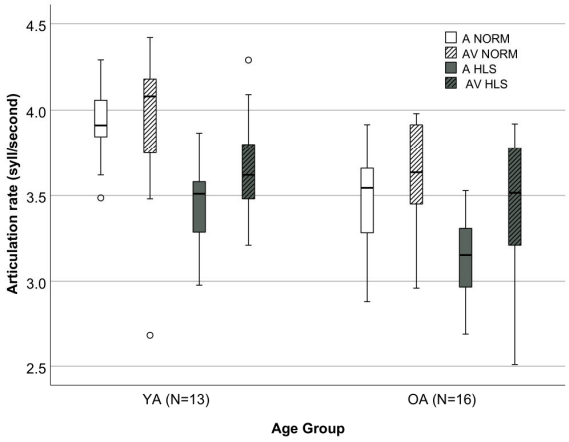
For median F0, there was no main effect of mode ($p=.233$) or age group ($p=.840$) but there was a crossover interaction, $F(1,28)=9.41$, $p=.005$, $\eta p^2=.25$: YAs decreased their median F0 in AV ($M=187$ Hz) relative to A ($M=196$), as shown in a paired t-test ($p=.002$) whilst OAs did not change their median F0 across modes ($M=188$ in A vs $M=192$ in AV).

For F0 range, there was no main effect of mode ($p=.937$) or group*mode interaction ($p=.271$) but there was an effect of age group, $F(1,28)=18.59$, $p<.001$, $\eta p^2=.40$: OAs had a wider F0 range ($M=40.5$ Hz) than YAs ($M=29.4$).

For the ME1-3 kHz measure, as OANH participants differed from OAHL, statistics are reported for them treated as separate groups. There was no main effect of mode ($p=.129$) or group*mode interaction ($p=.554$) but there was an effect of group, $F(2,27)=4.41$, $p=.02$, $\eta p^2=.25$: there was less mid-frequency energy in the voice of OAHL ($M=62.8$) than that of YA ($M=66.0$) and OANH ($M=65.1$) talkers who did not differ.

In summary, seeing their conversational partner when in good listening conditions had no effect on articulation rate, F0 range or on the ME1-3 kHz measure. The only effect was on median F0, which was lower in AV for YAs but not OAs. Overall, OAs did not show a different pattern of behaviour to YAs.

Figure 1: Articulation rate in syllables per second in both presentation modes and communicative conditions for the YA and OA groups



3.3. Clear speech adaptations in HLS in A mode

In [10], the analysis of the full corpus in A mode showed that talkers produced clear speech adaptations in the HLS condition. We first checked whether this was the case for this talker subset. Articulation rate (Fig. 1) was slower in HLS than in NORM, $F(1,28)=114.30$, $p<.001$, $\eta p^2=.80$. Also, talkers spoke with a higher median F0, $F(1,28)=16.54$, $p<.001$, $\eta p^2=.37$; a wider F0 range, $F(1,28)=18.31$, $p<.001$, $\eta p^2=.39$; and higher mid-frequency energy, $F(1,28)=47.32$, $p<.001$, $\eta p^2=.63$ than in NORM. Group effects were significant for articulation rate, F0 range and mid-frequency energy:

YAs spoke faster with a narrower F0 range and more mid-frequency energy. For this talker sample, the lack of group*condition interactions suggested that OA talkers made similar adaptations to YA talkers. However, as in [10], only the OAHL group showed correlated increases in median F0 and mid-frequency energy, considered a marker of vocal effort (Table 2).

Table 2: Correlation between percent relative increase in mid-frequency energy and in median F0 in NORM relative to HLS conditions (*=p<.05, **=p<.01).

	A	AV
YA (N=13)	-.578*	-.074
OANH (N=9)	.126	.646
OAHL (N=8)	.936**	.835**

3.4. Effect of visual cues on suprasegmental characteristics in HLS condition

This analysis investigated the impact of seeing the 'impaired' partner on clear speech adaptations (comparison of HLS in A vs HLS in AV).

Articulation rate (Fig.1) was faster in AV (M=3.56) than in A (M=3.30), $F(1,28)=24.32$, $p<.001$, $\eta p^2= .46$ for both age groups (mode*group, p=.613). YAs (M=3.56) spoke faster than OAs (M=3.30), $F(1,28)=6.08$, p=.02, $\eta p^2= .18$.

For median F0, neither effects of mode (p=.084), group (p=.942) or mode*group (p=.389) were significant. For F0 range, neither the effect of mode (p=.245), nor mode*group (p=.798) were significant. OAs had a wider F0 range (M=46.6) than YAs (M=35.9), $F(1,28)=16.91$, $p<.001$, $\eta p^2= .38$.

For ME1-3 kHz, as OANH participants differed from OAHL, statistics are reported for them as separate groups. ME1-3 kHz was higher in A (M=66.5) than AV (M=65.9), $F(1,27)=4.63$, $p<.05$, $\eta p^2= .15$. Mode*group was not significant (p=.080). ME1-3 kHz was lower for OAHL (M=64.8) than YA (M=67.5) and OANH (M=66.3) talkers who did not differ, $F(2,27)=3.43$, $p<.05$, $\eta p^2= .20$. In AV, again only the OAHL group showed correlated increases in median F0 and mid-frequency energy, considered a marker of vocal effort (Table 2).

In summary, when talkers could see their partner with 'impaired' hearing, they spoke more quickly and with less mid-frequency energy than when they could not see them, but pitch median or range did not change. As in A mode, OA and YA groups showed similar patterns of modifications in HLS relative to NORM when they could see their interlocutor. The correlation between increases in median F0 and ME1-3 kHz for the OAHL group suggests that face-to-face communication did not eliminate the increase in vocal effort that was present in the A mode for this group.

3.5. Relation between gaze frequency and acoustic measures of speech production

A significant level of p=.01 was used due to multiple comparisons. No significant correlations were obtained between gaze count in NORM and any of the acoustic measures. In HLS, there was a trend for gaze count to be correlated with the percentage of relative change in ME1-3 kHz from A to AV in HLS ($r(129)=.465$, p=.011): talkers who looked frequently at their interlocutor also showed a greater increase in mid-frequency energy in their voice in AV.

4. DISCUSSION

We investigated how seeing a conversational partner affected some aspects of speech production in easy and difficult communicative conditions. Seeing one's partner had little effect on suprasegmental features when communication was easy. This was also the case for older talkers. In HLS, both OA and YA participants looked more frequently at their 'impaired' partner to facilitate communication, but individuals in both age groups varied widely in gaze frequency and duration. It should be noted that the diapix task did require participants to look down at their pictures, which may have affected gaze frequency and duration, but this was the case for all participants. Contrary to findings in [9], face-to-face communication led to a reduction in some suprasegmental speech adaptations although the OAHL group still produced speech with increased vocal effort. This supports the view that speaker effort is adapted to interlocutor needs [11] and will be reduced when the interlocutor is provided with visual cues as an aid to communication [8]. The lack of a strong correlation between gaze frequency and acoustic measures may be due to different strategies being at play. Whilst gazes to Talker B may increase communication efficiency and reduce the need for acoustic adaptations, instances of miscomprehension by Talker B may lead individuals to both look at their interlocutor and increase their acoustic adaptations to enable effective communication to be re-established. A more detailed analysis of the interactions is necessary to provide a more fine-grained account of the effect of visual cues.

5. ACKNOWLEDGEMENTS

This work was supported by the Economic and Social Research Council [grant number ES/L007002/1]. We thank Leo Chong for his assistance in processing the video data.

6. REFERENCES

[1] ANVIL video annotation software. http://www.anvil-software.org/

[2] Baker, R., Hazan, V. 2011. DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods* 43, 761-770.

[3] Boersma, P., Weenink, D. 2018. Praat: doing phonetics by computer. http://www.praat.org/

[4] Cloud transcription service. https://www.speechmatics.com

[5] Cooke, M., King, S., Garnier, M., Aubanel, V. 2014. The listening talker: A review of human and algorithmic context-induced modifications of speech. *Computer Speech and Language* 28, 543-571.

[6] Davis, C., Kim, J., Tuomainen, O., Hazan, V. 2017. The effect of age and hearing loss on partner-directed gaze in a communicative task. *Proc. 14th AVSP Stockholm*, D1.S1.2 [online].

[7] De Looze, C., Hirst, D. J. 2008. Detecting key and range for the automatic modelling and coding of intonation, *Actes de Speech Prosody 2008 Conference, Campinas*, 135-138.

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

[9] Hazan, V., Kim, J. 2013. Acoustic and visual adaptations in speech produced to counter adverse listening conditions. *Proc. 12th AVSP Annecy*, 93-98.

[10] Hazan, V. L., Tuomainen, O., Kim, J., Davis, C., Sheffield, B., Brungart, D. 2018. Clear speech adaptations in spontaneous speech produced by young and older adults. *Journal of the Acoustical Society of America* 144, 1331-1346.

[11] Lindblom, B. 1990. Explaining phonetic variation: A sketch of the H&H theory. In: Hardcastle, W.J. & Marchal, A. (eds.), *Speech production and speech modelling* Dordrecht, the Netherlands: Kluwer Academic, 403–439.

[12] Mattys, S. L., Davis, M. H., Bradlow, A. R., Scott, S. K. 2012. Speech recognition in adverse conditions: A review. *Language and Cognitive Processes* 27, 953-978.

[13] Rinker, T. W. 2013. qdap: Quantitative discourse analysis package. version 1.3.1. Buffalo, NY: University at Buffalo.http://github.com/trinker/qdap

[14] Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., Bradlow, A. R. 2010. The Wildcat Corpus of Native- and Foreign-Accented English: communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech* 53, 510-540.