**elderLUCID CORPUS**

**RECORDING AND ANALYSIS METHODS**

## A. PARTICIPANTS

A total of 98 speakers were recruited for this study. Data from seven speakers were excluded because of their language background, age or hearing level, a further three volunteers withdrew from the study after the first screening session and five volunteers did not manage to finish the three sessions on time. The volunteers were recruited via UCL Psychology subject pool, via University of the 3rd Age (London) and greater London rambling groups. The final sample consisted of 83 single-sex pairs of native Southern British English adult talkers between the ages of 19 and 84 years. The pairs did not know each other and had not met each other before the recordings. Each participant was assigned a role of a primary or the lead talker ('Talker A') or a secondary talker ('Talker B'). Primary talkers were divided into two age groups: 'younger adults' (YA) between 19-26 years of age (15 F, 11 M; F Mean=22 years, M Mean=21 years) and 'older adults' (OA) between 65-84 years of age (30 F, 27 M; F Mean=71 years, M Mean=74 years). Secondary talkers were always younger adults (N=83, between 18-30 years of age) of the same sex as the Talker A. Participants reported no history of speech or language impairments. YA participants passed a hearing screen at 25 dB HL or better at octave frequencies between 250-8000 Hz in both ears. OA participants had either 'normal hearing' (OANH: Female N=14; Male N=13) defined as a hearing threshold of <20 dB between octave frequencies 250-4000 Hz or a 'mild hearing loss' (OAHL: Female N=16; Male N=14) defined as a hearing threshold of <45 dB in this frequency range with a symmetrical downward slope of pure tone threshold in the high-frequency range typical for an age-related hearing loss profile. Informed written consent was obtained (also for the recordings). Ethical approval was obtained from the University College London (UCL) Research Ethics Committee.

## B. DIAPIX AND BKB TASKS

Spontaneous speech dialogs were elicited using the diapix task (van Engen et al., 2010), a 'spot the difference' picture task. Use was made of the diapixUK picture materials developed by Baker and Hazan (2011). The pictures included hand-drawn scenes produced by an artist, which were then coloured in; these were designed to be fairly humorous to maintain interest in the task. Each picture included different 'mini-scenes' in the four quadrants of the picture, and the differences to be found were fairly evenly distributed across the four quadrants. These differences could be differences in an object or action across the two pictures (e.g., green ball in picture 1 vs red ball in picture 2; holding the ball in picture 1 vs kicking the ball in picture 2) or omissions in one of the pictures (e.g., missing object on a table in one picture). All four versions of the 'beach' scenes, 'farm' scenes and 'street' scenes of the diapixUK picture set were used, and pictures were counterbalanced across talker pairs.

The full set of pictures is available as supplementary materials to the following article and downloadable from the website below:
Baker, R., Hazan, V., 2011. DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. Behav. Res. Methods 43, 761-770.
http://link.springer.com/article/10.3758%2Fs13428-011-0075-y

The BKB task was designed to be an interactive sentence-reading task. In this task, the Talker A read BKB sentences (Bench & Bamford, 1979) one by one to Talker B whose task was to listen and repeat the sentence. The BKB sentences are short with a simple syntax and they are meaningful and refer to everyday situations, such as 'The boy followed the path' and 'The postman shut the gate'. The talker pairs were told that their performance was scored based on how many words correct they get.

These two tasks were chosen because the Diapix and the BKB tasks produce sentences that are similar in complexity, but reading sentences requires less planning and cognitive resources than producing spontaneous speech within a problem-solving task. Both BKB and Diapix tasks were run in a condition where the two speakers could see each other (Audio-Visual, AV, data not included here) and when they could not see each other (i.e., they could only use auditory information to solve the task, Audio-only, A). The two conditions were run on separate days and their order was randomised within the age groups.

### C. RECORDING PROCEDURE

Lead talkers (Talker A) visited the lab for three 2-hour sessions: a 2-hour initial screening session where various sensory (PTA, Frequency modulation detection, gap duration detection, speech-in-noise test) and cognitive tasks (short-term memory, working memory, word association) were run. After this, they were invited back within 2 days to 2 months to complete the two communicative sessions with the same-sex secondary talker. Each pair of participants was tested over two two-hour sessions during which they completed eight Diapix tasks and eight BKB tasks in different environmental conditions and when they could see each other (Audio-Visual condition, not included here) and when they could not see each other (Audio condition). The various conditions of the diapix tasks were distributed across the two sessions and the order of presentation of the conditions was counterbalanced across participant pairs.

During the recording, the two participants sat in different rooms and communicated via Vic Firth headsets fitted with an Eagle G157b lapel microphones. The speech of each participant was recorded on a separate channel at a sampling rate of 44,100 Hz (16 bit) using an EMU 0404 USB audio interface and Adobe Audition and Rode NT1-A condenser microphones. In the ‘normal’ (NORM) condition, the two speakers could hear each other without difficulty. In order to elicit clear speech adaptations, in the hearing loss simulation condition (HLS) the voice of one of the talkers ('Talker A') was processed in real time through a hearing loss simulator (HELPS; Zurek, and Desloge, 2007) mimicking the effect of severe-to-profound age-related hearing loss before being transmitted to Talker B. The simulation therefore affected speech intelligibility for Talker B and it was expected that, in this condition speaker A would have to clarify his/her speech for the benefit of talker B. In the BABBLE (BAB-1) condition, the speech of talker A was mixed with the same 8-talker babble as used in the earlier adult diapix study (Hazan and Baker, 2011) before being channelled through to the confederate’s headphones, at a difficulty level equated to the HLS conditions via a Modified Rhyme Task (MRT). In the second BABBLE (BAB-2) conditions, both talkers, A and B, heard the same background babble as in BAB-1 but now at an approximate level of 0 dB SNR.

To familiarise participants with the roles of talker A and B and the nature of differences typically found in the picture sets, participants began by receiving training on the diapix task with a set of pictures that was never used in the recordings. They were each given a picture and sat so that they could not see each other's picture. They were told the pictures contained 12 differences which they had to find. Recordings were stopped once all differences had been found or after ten minutes. Talker A ('the leader') was instructed to do most of the talking and reading, whereas the Talker B was mainly there to ask questions and make suggestions (diapix) or repeat back the sentences (BKB). In the diapix task, the participants were told they had 10 minutes to find the 12 differences. They were told that during some of the recordings the voice of speaker A would be distorted, and that the experimenter would inform them when this was about to happen. In the BKB task, the participants were told to read and repeat all 32 sentences per listening condition, and that their aim is to score as high a score as possible.

Every pair of participants carried out eight recordings with different sets of pictures: two without a communication barrier (NORM), two in the HLS condition and four in babble (BAB-1, BAB-2).

Every pair started out with a recording in NORM, and pairs of participants were counterbalanced between doing HLS and two babble conditions.

The full set of recordings will be available online within the UCL data haven archive.

## D. BACKGROUND COGNITIVE AND SENSORY FUNCTION SCREENING TASKS

In order to investigate the relationship between sensory and cognitive function and speech adaptations, a test battery of measures was carried out on the primary participants.

### a) Sensory function.

To obtain hearing thresholds, pure tone audiograms were measured in both ears at octave frequencies between 250-8000 Hz (BSA, 2011). Psychophysical tests of gap and frequency modulation were run using adaptative procedures described in Schoof and Rosen (2014).

Thresholds for word-intelligibility in background noise were obtained using the task as described in Hazan et al., 2009 (WiNics), a task modelled on the coordinate response measure (Moore, 1981). In summary, participants heard the following carrier phrase: "Show the dog where the [color] [number] is," using 6 colours (black, red, white, blue, green and pink) and 8 digits (1,2,3,4,5,6,8,9). In a trial, participants had to click on the symbol showing the colour/number combination. A three-up/one-down adaptive procedure was used to vary SNR and to track 79.4% correct trials. The threshold for a 79.4% correct level was calculated from the mean of the reversals excluding the first two.

### b) Cognitive function.

To obtain measures of short-term memory, a digit span forward (DSF) task was used. For working memory, we used the backward digit span (DSB) test that measures information storage and rehearsal. In the DSF/DSB tasks, the participant repeated auditory presented number sequences in the same or reverse order and they were scored correct/incorrect for each sequence (maximum scores, DSF=16, DSB=14).The efficiency of word search and retrieval from the stored lexicon was measured using the Word Association (WA) task in which participants had to say as many words as possible from a category in 60 seconds (final score is a total number of items across the three categories). All tasks from: Semel, E., Wiig, E. H., & Secord, W. A. (2006).

## E. DATA PRE-PROCESSING

For all recordings, each channel was automatically transcribed using cloud-based speech recognition system by Speechmatics (https://www.speechmatics.com/). An in-house Matlab script was used to upload and download files to the Speechmatics database and create a Praat textgrid from the original JSON format. These automated transcriptions and the audio-transcription alignment were then hand-checked at a word level and corrected for errors to a set of transcription guidelines based on those used by Van Engen et al. (2010) with minor adaptations for the coding of pauses.

Phoneme-level alignment software that was developed in-house at UCL was used to automatically align the transcriptions and create Praat Textgrids with separate phoneme tier in addition to the word tier. Recordings lasted for about 10 minutes, yielding around 4 minutes of analysable speech for Talker A once silences, fillers, non-speech sounds such as laughter and sections with background noise had been excluded.

The Praat textgrids containing alignments at word and phoneme level will be available for download on request to the corpus authors.

## F. DATA ANALYSIS

### a) Measures of communication efficiency

In the Diapix task, the measure taken as reflecting communication efficiency was the time taken to find the first eight differences (time8) in the picture 'spot the difference' task in each condition. This criterion was chosen as the threshold as not all participants found all 12 differences in the picture. The number of differences found before the task was terminated (ten minutes or when all differences had been found) is also given. In the BKB task, the measure taken to reflect communication efficiency was the percentage keywords correct in each listening condition.

## b) Acoustic analyses

### *Fundamental frequency median and range*

Fundamental frequency analyses were carried out using Praat software on each of the recordings for Talker A in each condition. A Praat script opened each file, extracted the intervals which were not marked as silences, laughter, noise or breath intake, and concatenated the extracted intervals. Then, on the concatenated file, f0 calculations were done using the ‘pitch’ function in Praat, with a time step of 100 pitch values per second. A formula as described in De Looze and Hirst (2008) was used to calculate ceiling and floor limits specific to each speaker rather than default values for males and females, as this has been shown to be more successful in excluding rogue values. For each speaker in each condition, long term distributional measures, i.e. median fundamental frequency and interquartile range (75th -25th percentile, i.e. 50% span), were calculated over the aggregated speech; semitones relative to 1 Hz were used as a measure to facilitate comparisons of pitch range across male and female talkers.

### *Intensity measures*

Long-term average spectrum (LTAS) analyses were carried out using a Praat script. First, for each file, the intensity of all labelled speech segments was calculated and those above a set level (88 dB) were excluded for the LTAS calculations, as likely to be instances of shouting. The remaining speech segments were concatenated and the intensity of the resulting waveform scaled to a set level (75 dB). The signal was then band-passed filtered between 1 and 3 kHz and the mean intensity of the resulting waveform calculated to give a measure of mean energy between 1 and 3 kHz, relative to the total energy in the spectrum. This measure was labelled as ‘Mean energy 1-3 kHz’. This frequency band was chosen because it has been shown that talkers increase their energy in this region when producing clear speech.

### *Articulation rate and pausing*

Articulation rate was calculated as the number of syllables produced by talker A divided by the total duration of speech segments (excluding fillers, silences, etc) for that talker. Syllable counts were calculated from the orthographic transcriptions of the spontaneous speech using the qdap package in R (Rinker, 2013). Segments labelled as unfinished words, hesitations, fillers and agreements (e.g. ‘yeah’, ‘yup’, ‘err’, ‘hmm’) were excluded from the speech duration analysis. In the same Praat script, a count was kept of the silent pauses that were longer than 300 ms, and their mean duration was also calculated.

### *Vowel measures*

Vowel area was examined by analysing three vowels in content words: [i:], [ae] and [o:]. These were chosen amongst available monophthongs because 1) they were the most frequent per individual participant recordings 2) they had the best differentiation in terms of front-back and high-low distinctions and therefore covered the largest distances in the F1-F2 quadrilateral space. These vowels were selected from content words produced in the spontaneous speech.

Formant estimates were normalized to ERB values to reduce the effect of anatomical differences due to gender and age, and median F1/F2 ERB values were calculated per vowel per talker.

For each talker, a measure of F1 range (in ERB) was derived by subtracting F1[i:] from F1[æ] , giving an indication of how much vowels were differentiated in terms of height. The degree to which the front/back distinction was instantiated was explored by examining the F2 range obtained by subtracting F2[i:] from F2[o:].

The acoustic vowel space was derived for individual speakers from F1-F2 values of the three vowels ([i:], [ae],[o:]) separately per condition. We first derived the Euclidean distance between pairs of vowels. Heron's formula was then used to calculate the vowel space between the three vowels.

**References**


Baker, R., Hazan, V., 2011. DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. Behav. Res. Methods 43, 761-770.

Bench, J,. Kowal, Å ., Bamford, J., 1979. The BKB (Bamford–Kowal–Bench) sentence lists for partially-hearing children, British. Journal of Audiology, vol. 13, pp. 108–112.

De Looze, C., Hirst, DJ., 2008. Detecting Key and Range for the Automatic Modelling and Coding of Intonation, Actes de Speech Prosody 2008 Conference, Campinas, Brazil, 135-138.

Hazan, V. L., Baker, R., 2011. Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. J. Acoust. Soc. Am. 130, 2139-2152.

Hazan, V., Messaoud-Galusi, S., Rosen, S., Nouwens, S., Shakespeare, B., 2009. Speech perception abilities of adults with dyslexia: is there any evidence for a true deficit?. Journal of Speech, Hearing and Language Research, 52(6), 1510-1529.

Rinker, T. W., 2013. qdap: Quantitative Discourse Analysis Package. version 1.3.1. University at Buffalo. Buffalo, New York. http://github.com/trinker/qdap

Semel, E., Wiig, E., and Secord, S., 2003. Clinical Evaluation of Language Fundamentals-4, The Psychological Corporation, San Antonio, Tex, USA.

Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M. and Bradlow, A. R., 2010. The Wildcat corpus of native- and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. Lang. Speech 53, 510-540.

Zurek, PM., Desloge, JG., 2007. Hearing loss and prosthesis simulation in audiology, Hearing Journal, Vol. 60, pp 32–33, 36,38. 2007.