kidLUCID CORPUS

RECORDING AND ANALYSIS METHODS

A. Participants

Fifty single-sex pairs of children aged 9 to 14 years inclusive were recruited for this study. The two talkers in each pair knew each other and were friends. Data from two male pairs could not be included because of non-completion of the recording sessions, resulting in a total of 96 child participants (46M, 50F, mean age: 11;8 years, range 9;0 to 15;0 years). Participants were native southern British English speakers who reported no history of hearing or language impairments. All participants passed a hearing screen at 25 dB HL or better at octave frequencies between 250 and 8000 Hz in both ears.

B. Diapix task

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 colored in; these were designed to be fairly humorous to maintain interest in the task (see Figure 1 for an example of one of the picture pairs). 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). The first three '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

C. Procedure

Each pair of participants was tested over two two-hour sessions during which they completed a range of tasks including: several diapix tasks in different environmental conditions, a picture naming task (not included here), pure tone hearing screening and an adaptive test of word perception in noise (not included here). At each session, one block of the picture naming task was carried out first, and the hearing screening was also carried out before the rest of the tests in the first session. 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 headsets fitted with a condenser cardioid microphone (Beyerdynamic DT297). 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. In the ('no barrier' NB) condition, the two speakers could hear each other without difficulty. In order to elicit clear

speech adaptations, in the vocoder condition (VOC) the voice of one of the talkers ('talker A') was processed in real time through a three-channel noise-excited vocoder before being transmitted to Talker B. The vocoded speech therefore affected speech intelligibility for Talker B and it was expected that, in this condition (VOC) speaker A would have to clarify his/her speech for the benefit of talker B. Given the significant learning effect when listening to vocoded speech and children's lack of familiarity with vocoded speech, all participants carried out a ten-minute computer-based training session with vocoded speech before taking part in the VOC condition. In the BABBLE condition, the speech of talker A was mixed with the same 8-talker babble as used in the adult diapix study (Hazan and Baker, 2011) before being channelled through to the confederate's headphones, at an approximate level of 0 dB SNR. The NH talker was told that their interlocutor would hear their speech in a background of lots of voices mixed together which would be quite loud compared to their voice.

To familiarise participants with the roles of talker A and B and the nature of differences typically found in the picture sets, children 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 the other's. 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. One child was designated 'the leader' (speaker A) and instructed to do most of the talking, whereas the other child (speaker B) was mainly there to ask questions and make suggestions. With the younger age group, the experimenter gave some of the more reticent participants hints during the training phase. After they had found several differences, children were allowed to look at each other's pictures and continue comparing them. The participants were told they would take turns at each role in separate recordings, and that they had 10 minutes to find the 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.

Every pair of participants carried out six recordings with different sets of pictures: two without a communication barrier (NB), two in the VOC condition and two in the BAB condition.

Every pair started out with a recording in NB, and pairs of participants were counterbalanced between doing two VOC recordings first or the BAB condition. Everyone ended with the second NB recording. Participants switched roles between recordings in each condition, so that each participant was recorded both as 'speaker A' (leading the interactions) and 'speaker B'.

The full set of recordings is available online within the OSCAAR archive (https://oscaar.ci.northwestern.edu/).

D. Data processing

For all recordings each channel was transcribed using freeware transcription software from Northwestern University's Linguistics Department (Wavescroller) to a set of transcription guidelines based on those used by Van Engen et al. (2010) with minor adaptations for the coding of pauses. Word- and phoneme-level alignment software that was developed in-house at UCL was used to automatically align the transcriptions and create Praat Textgrids with separate word and phoneme tiers. Alignment was manually checked and corrected in two stages: first the word level alignment of all the files was manually checked and adjusted where necessary. Corrected word level textgrids were automatically re-aligned to correct the phoneme level. The alignment of the three vowels that were analysed for the vowel space

measures was then verified and corrected by hand where necessary in these new files. 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 work and phoneme level are available for download from the kidLUCID corpus section of the OSCAAR archive (https://oscaar.ci.northwestern.edu/). Transcripts of the interactions (with time stamps) are available for download from a link on our project website (http://www.ucl.ac.uk/pals/research/shaps/research/shaps/research/clear-speech-strategies).

MEASURES OF COMMUNICATION EFFICIENCY

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. Another measure of communication efficiency, not dependent on speaking rate, is the total number of words produced by Speaker A (instructed to take the lead in the task) to task completion in each condition (word_count). Communication between the two talkers was judged to be maximally efficient when they managed to complete the spot-the-difference task with the least number of words produced by the talker leading the interaction.

CLARITY RATINGS MEASURES

Four short samples were extracted for each talker from each of the three conditions. Similar to the study by Hazan & Baker (2011), these samples were extracted from as close as possible to the 10th, 15th, 20th and 25th turn in each conversation. Each sample was between 2-3 seconds long, and they were either a whole intonational phrase or the end of a phrase and did not occur after a miscommunication (e.g., a clarification request). Twenty-four native Southern British English speakers all with normal hearing (5M, 19F; mean age: 24;3, range: 19;4-31;0 years) took part in the rating experiment. The randomised speech samples were presented via headphones (Beyerdynamic DT297) in two separate sessions run with a minimum of 2 hours between the sessions. The listeners rated the clarity of each speech sample on a scale 1-7 ("1" clear and "7" not very clear). Mean ratings were calculated for each speaker per condition.

ACOUSTIC-PHONETIC MEASURES

A number of acoustic-phonetic measures were selected; these were the features that were also analysed for our adult diapix corpus (Hazan and Baker, 2011). These include measures of fundamental frequency median and range, mean word duration (reflecting speech rate), mean energy in the 1-3 kHz range of the long-term average spectrum of speech, and vowel space. These measures were carried out both on the diapix.

1. Fundamental frequency median and range

Fundamental frequency analyses were done in Praat on each of the recordings for Talker A in each condition. A Praat script opened each file, extracted the intervals which are not blank or marked as silences, laughter, noise or breath intake, and concatenated the extracted intervals. Then, on the concatenated file, the pitch extraction was calculated using the 'pitch' function in Praat, using a time step of 150 pitch values per second, a pitch floor value of 50Hz and

pitch ceiling value of 500 Hz. Median fundamental frequency and interquartile range (difference between 1st and 3rd quantile) were then calculated for each file in semitones re 1 Hz. A median value was preferred to the mean to reduce the effect of inaccurate period calculations, which are likely in spontaneous speech, while semitones were used to facilitate comparisons across male and female talkers.

2. Intensity measures

Long-term average spectrum (LTAS) analyses were also 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 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. The signal was then band-passed filtered and the mean intensity of the resulting waveform calculated to give the measure of mean energy between 1 and 3 kHz.

3. Articulation rate

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 500 ms, and their mean duration was also calculated.

4. Vowel measures

Vowel area was examined by analysing three corner 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.

On average, 29 [i:], 21 [æ] and 15 [o:] vowel tokens were included in the calculations of vowel measures per talker for NB and 25 [i:], 18 [æ] and 14 [o:] tokens for the vowel tokens were included in the calculations of vowel measures per talker for NB and VOC conditions, respectively. 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 speaker, 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[ao].

The acoustic vowel space was derived for individual speakers from F1-F2 values of the three vowels ([i], [ae],[ao]) 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.

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.

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

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.