Proceedings of DiSS-LPSS Joint Workshop 2010,
25-26 September 2010, Tokyo, Japan, pp. 03-06.

LUCID: a corpus of spontaneous and read clear speech in British English

Rachel Baker[1], *Valerie Hazan* [2]

Department of Speech, Hearing and Phonetic Sciences, UCL, UK

rachel.baker@ucl.ac.uk[1], v.hazan@ucl.ac.uk[2]

Abstract

This paper describes LUCID, the London UCL Clear Speech in Interaction Database, which contains spontaneous and read speech in clear and casual speaking styles for 40 Southern British English speakers. The problem-solving task used to collect the spontaneous speech, the DiapixUK task, is also described, along with ways of using the task to elicit different types of clear speech without explicit instruction, e,g. using different 'barriers' to communication. Applications of the corpus and of the task materials for future research projects are discussed. The corpus and materials will be available online to the research community at the end of the project.

Index Terms: spontaneous speech, speech production, clear speech, interaction

1. Introduction

This paper describes LUCID, a corpus which was primarily created to investigate clear speech produced with and without communicative intent, i.e. in spontaneous and read speech. Most previous studies of clear speech production have been based on recordings of read speech, with the speaking styles elicited through different sets of instructions (e.g. [1, 2]). Compared to read speech produced casually (e.g. as if talking to a friend) read speech which is produced clearly (e.g. as if talking to someone who is hearing impaired) has a higher mean and greater range in fundamental frequency, a slower speech rate and a more expanded vowel space, amongst other differences (see [3] for a review of clear speech research). However, read speech has different phonetic characteristics to spontaneous speech (e.g. [4]) which leads to the following questions. Is read clear speech representative of spontaneous speech which is clarified to facilitate communication in adverse listening conditions? Does clear speech vary according to the type of communication barrier encountered? We present some evidence relating to these questions in [5] and focus in this paper on the design of the task used to elicit our corpus and on a description of the LUCID corpus itself.

1.1. The need for a new spontaneous speech task

A number of tasks have previously been used to elicit spontaneous speech for research purposes. The Map Task [6] is probably the best-known of these; it involves an 'instruction giver' communicating details of a map route to an 'instruction follower' who has no indication of the route on their map and different key elements on the map itself. The Map Task has been used in a number of studies to elicit spontaneous dialogues with a constrained topic and/or specific keywords. Studies have also used other problem-solving tasks to record spontaneous speech produced by two participants, e.g. [7] recorded two pairs of people in the same room, each solving a different crossword in order to record spontaneous speech produced when there are competing talkers present. More recently, [8] asked participants to try to complete Sudoku puzzles in order to obtain many repetitions of number words.

Recently, [9] developed the 'Diapix' task, which is also a problem-solving task involving two talkers. Each talker is presented with a different version of the same cartoon-style picture and the two talkers have to collaborate to find ten differences between the two pictures. Specific keywords can be elicited via the differences that need to be found; generally, the content of the speech is quite similar across talkers and conditions due to the constrained topic of conversation. The Diapix task shares many similarities with the Map Task but offers added flexibility in terms of the relationship between participants. Participants have equal roles if both are instructed to contribute to finding the differences. Alternatively, one participant can be instructed to 'take the lead' in the task, resulting in an information giver/receiver relationship.

Using the Diapix task, the communicative situation can be controlled to naturally elicit clear speech in one or both talkers. It would presumably be possible to do this using the Map Task, crosswords, and Sudoku puzzles but to our knowledge, those tasks were not used in that way. [9] controlled communication difficulty by comparing the speech produced by a talker when communicating with two different partners: a talker with a shared native language and a non-native talker. Simple measures such as the time taken to complete the task or the balance of speech between the two talkers were sensitive enough to show differences in communication difficulty between native English speaker pairs, native–non-native pairs and between non-native pairs of matched or unmatched L1s. The Diapix task is therefore a promising technique for further investigations of clear speech elicited in a dialogue situation.

The original Diapix pictures (3 pairs) were not suitable for our study because our design required each pair of talkers to repeat the task in many different conditions, so that a large set of picture-pairs of similar difficulty was needed. The original Diapix pictures had been hand-drawn and saved in hard copy so were not able to be modified or extended to suit our research goals. Therefore a new version of the Diapix task, DiapixUK, was created. This contains a set of 12 picture pairs (plus a training picture), which were used in our study to elicit casual and clear speech in different 'adverse communication' conditions.

2. Speech corpus design and collection

2.1. Participants

Forty native speakers of Southern British English (20 M, 20 F; 19 - 29 yrs old) served as main participants. They were all students or staff from the University of London. Participants volunteered with a friend of the same gender so there were 20 'friend' pairs (10 M, 10 F) although mixed gender pairs could also have been used. Participants read 'accent-revealing' sentences before being accepted onto the study, to ensure that they were from the appropriate accent group. Another eight native (N) speakers of Southern British English (4 M, 4 F) who fitted the above criteria were recruited as confederates for the recording session involving background noise.

Six non-native (NN) speakers of English (23-39 yrs old) were recruited as confederates for the recording session involving N-NN English conversations. The NN English speaker group comprised: two Mandarin Chinese speakers (1 M, 1 F), two female Taiwanese Chinese speakers and two male Korean speakers. They were selected to be similar in their degree of oral proficiency as judged by their performance on a standardized test of English language skills which involved an oral proficiency component (Versant). They all obtained a Versant English test (overall) score within the 4th and 5th ability groups (1st group: highest ability; 6th group: lowest ability).

All participants had normal hearing thresholds (20dB HL or better for the range 250 – 8000Hz) and reported no history of speech or language disorders. All but two of the main participants had no specific experience communicating with people with speech and language difficulties. Participants were naive of the purpose of the recordings but were debriefed afterwards and paid for their participation.

2.2. Materials

2.2.1. DiapixUK task – spontaneous speech

Each DiapixUK task consists of two versions of the same cartoon picture which contain twelve differences. Each person is given a different version of the picture and is seated in a separate sound-treated room (without view of the other person). The two speakers communicate via headsets to locate the twelve differences between the two pictures.

Twelve picture-pairs were created which belong to one of three themes: beach scenes (B), farm scenes (F), and street scenes (S) with four pairs per theme. Figure 1 shows farm scene 3. The pictures were based on hand-drawn scenes produced by an artist; these were then scanned to create digital line drawings which were coloured in using Adobe Photoshop. Each item in the pictures, e.g. an object, a person etc, was assigned to a separate layer in Photoshop to allow for easy modification of the pictures, if this were to be needed in other studies. The scenes were designed to be fairly humorous to maintain interest in the task and were drawn in a cartoon style. Each picture includes a different 'mini-scene' in each of the four quadrants, and the twelve differences were fairly evenly distributed across these scenes. Each difference was designed to encourage elicitation of one of 36 keywords. Each keyword is a monosyllabic CV(C) word that belongs to a (near) minimal word pair with the /p/-/b/ or /s/-/ʃ/ contrasts in initial position (e.g., *pear*/*bear*; *sign*/*shine*). This allows for the analysis of the production of these two contrasts in clear and casual spontaneous speech. The differences are either differences in an object or action across the two pictures (e.g. red ball of wool in picture A vs. blue ball in picture B; holding a beach ball in picture A vs. sitting on a beach ball in picture B) or omissions in one of the pictures (e.g. sign on a shop in picture A vs. no sign in picture B). Some of the differences are visual while others consist of differences in writing (e.g. on posters or shop fronts) across the two pictures. The 36 keywords were divided into three sets and each set of twelve was used for a different picture theme, i.e. beach, farm or street. As a result, completion of three diapix tasks: 1 beach, 1 street and 1 farm scene, would be likely to result in the production of all 36 keywords. It is important to note that although the diapix pictures were designed to elicit specific keywords, global acoustic-phonetic measures can also be carried out on the entirety of the speech produced by each talker, so the specific analysis of keywords is not essential.

A training picture pair was also created. It follows the same format as the other picture pairs except that the twelve differences are not related to the keyword set. This picture can be used to familiarize participants with the task procedure before any of the main picture pairs are used.

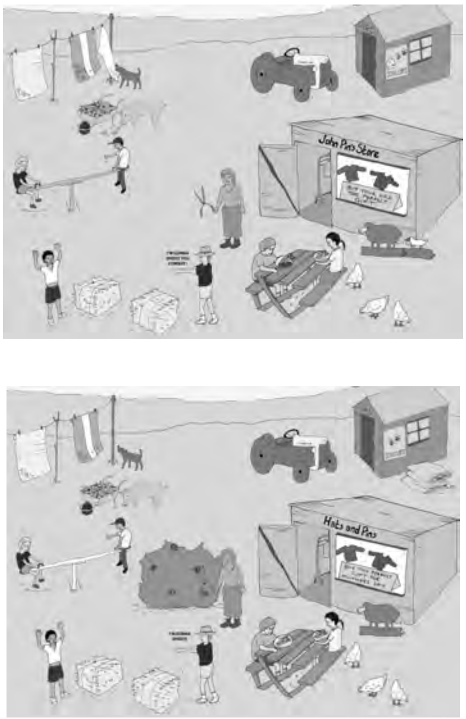

Figure 1: *Example of DiapixUK task picture pair: Farm scene 3 (top: version A; bottom: version B)*

During the task development, a pilot study was conducted to investigate three things: the possibility of a learning effect by participating in multiple diapix tasks, the comparative level of difficulty of each picture-pair and the likelihood of each keyword being produced. Twelve native British English participant pairs took part in the pilot study. Each pair began the study with the training task and then did 6 tasks (2 Beach, 2 Farm, 2 Street) in an order following a modified Latin square design. Transaction time, defined as the time taken to find 10 differences, was used to compare performance on each picture pair in each position. There was no significant difference in transaction time due to picture position [F(5,62)=1.2,p=0.33] suggesting that talkers did not get quicker at the task the more pictures they completed. The lack of a significant learning effect means that it is reasonable to ask participants to complete more than one picture pair. There was however, a main effect of picture number [F(3,64)=4.0,p=0.01] suggesting that not all of the picture pairs were of equal difficulty. Consequently, all picture pairs were reassessed. Since those with longer transaction times were the minority and to retain the pictures' suitability for younger age groups, those with significantly longer transaction times were simplified, using the recorded conversations as pointers to the parts of the pictures that resulted in unnecessarily detailed descriptions. The revised forms of these pictures were not piloted, but our careful modifications resulted in new versions that we believe to be of approximate equal difficulty. This is supported by a lack of difference between picture versions in the main corpus. Some differences were less likely to elicit a keyword than others. Where possible, differences were modified to further increase the chance of a keyword being produced.

2.2.2. *Picture naming and sentence reading tasks*

To collect multiple iterations of specific phonetic contrasts, participants also completed picture naming and sentence reading tasks. For the picture naming task, a recognizable picture of each of the 36 /p,b,s,ʃ/ keywords was found. 30 keywords were represented by a picture of a noun, e.g. *ball*, and 6 were represented by a picture of a verb, e.g. *push*. For the sentence reading task, 4 sentence pairs were created for each of the 18 keyword pairs (total: 144 sentences). Within each sentence pair, keywords were matched for prosodic position and preceding phonetic context, e.g. 'the old lady ate the *peach*/the young children loved the *beach*'. Keyword position in the sentence was varied across pairs.

2.3. Recording

Beyerdynamic DT297PV headsets were used in all sessions and speech was recorded at a sample rate of 44,100HZ (16 bit) using an E-MU 0404 USB audio interface and Adobe Audition or DMDX [10]. The speech of each talker was saved on a separate audio channel for transcription and acoustic analysis.

Each participant was involved in five recording sessions as detailed below. The order in which pictures were completed at each diapix session was counterbalanced across conditions and participant pairs following a modified Latin square design. In these sessions, participants were told to start the task in the top left corner of the picture and work in a clockwise manner around the scene. The experimenter monitored the recording from outside both of the recording rooms. She stopped each recording either once the 12 differences were found, or after at least 15 minutes had lapsed and participants could not locate the final differences.

Three different 'communication barriers', affecting one of the two participants only, were used in the diapix sessions to elicit clear speech in the participant hearing normally (see Sections 2.3.2 and 2.3.3). Different types of communication barriers were used to see whether the clear speech that was elicited in the participant hearing normally varied according to the type of degradation the other participant was experiencing.

2.3.1. *Session 1: spontaneous casual speech – Diapix No Barrier (NB)*

This session was completed in good listening conditions. Each pair did the training diapix task, then three diapix tasks (1 B, 1 F and 1 S). Both participants were asked to contribute to finding the differences to encourage a balanced conversation.

2.3.2. *Session 2: spontaneous clear speech – Diapix VOC*

In this condition, one of the talkers heard the speech of the other participant after it had been processed in real-time through a three-channel noise vocoder. The vocoder spectrally degrades speech, as the full spectrum is filtered through three filters only and pitch information is lost as the vocoder is noise-excited. A three channel vocoder introduced enough difficulty to the task to necessitate clear speech from the participant who was hearing normally while still allowing enough communication to do the task.

As there is a significant learning effect when listening to vocoded speech, immediately prior to the task itself, each participant completed a ten-minute 'vocoder-familiarisation' task. Each pair then completed six diapix tasks in total: three when the first participant's speech was vocoded and three when the other participant's speech was vocoded. The participant who was hearing normally was encouraged to take the lead in the conversation to discourage the 'impaired' participant, i.e. the person who was hearing the vocoded speech, from dominating the conversation.

2.3.3. *Session 3: spontaneous clear speech – Diapix BABBLE or L2*

In session 3, half of the participants did the task in the BABBLE condition, while the other half did the task in the L2 condition. In the BABBLE condition, 20 participants (11 M, 9 F) did three diapix tasks with an 'impaired' native confederate of the same gender who heard the speech of the other participant mixed in real-time with pre-recorded 8-talker babble. The confederate was familiar with the task procedure having previously done the training diapix task in normal listening conditions. In the L2 condition, the remaining participants (9 M, 11 F) did three diapix tasks with one of the NN confederates. The NN confederate had previously done the training diapix task as a means of familiarization with the task procedure.

2.3.4. *Sessions 4 & 5: Naming/read casual & clear speech*

Each participant completed two individual recording sessions with stimuli presented via DMDX [10]. In session 4, they first did the picture naming task. Participants were instructed to name each picture using one of two frame sentences: 'I can see a (noun)' or 'The verb is to (verb)'. The 36 pictures were presented 8 times in a pseudo-randomized order (nouns and verbs were separate). The sentence reading task followed. Participants read sentences from a computer screen in a pseudo-randomized order. For both tasks, participants were asked to speak 'casually as if talking to a friend'. In session 5, participants did the same tasks but were instructed to speak 'clearly as if talking to someone who is hearing impaired'. These are instructions typically given in studies of clear speech that have involved read speech materials.

2.4. Sound file post-processing

For all files, the speech of each participant (excluding NN confederates) was orthographically transcribed using freeware transcription software (Wavescroller) to guidelines used by [9]. The transcripts were automatically word-aligned to the sound files using NUAligner software, which created a Praat TextGrid. This alignment was hand-checked in approximately two-thirds of the file set. In each file, a second transcription tier was created in which all content words are phonemically-transcribed. All audio files were normalized to a mean amplitude of 15dB (with soft limiting) in Adobe Audition.

3. Applications of the corpus

The corpus contains approximately 110 hours of recordings of spontaneous and read speech in clear and casual styles. The mean duration of actual speech (i.e., linguistic material only, excluding silences and pauses) per participant per condition is: spontaneous casual speech: 2.6 minutes; spontaneous clear speech (VOC): 4 minutes; spontaneous clear speech (BABBLE): 4 minutes; spontaneous clear speech (L2): 5.3 minutes; read casual speech: 4.1 minutes; read clear speech: 6.5 minutes. The total number of words in the corpus is approximately 420,000.

A comparison of transaction time (time taken to find the first eight differences) across the casual and clear speech diapix conditions verified that the three 'communication barriers' created enough difficulty to disrupt communication in the tasks relative to the 'no barrier' (NB) control condition (see Table 1 for mean values). The transaction time was significantly longer for the VOC condition than the NB condition [F(1,36)=64.4; p<0.001]. In separate pairwise comparisons, transaction time was also significantly longer for the L2 condition than the NB condition (p<0.0001) and it was marginally significantly longer for the BABBLE condition than the NB condition (p<0.09). This analysis shows that the interactions were more effortful in the VOC, L2 and BABBLE than in the NB condition, which is associated with clearer speech being produced in the three barrier conditions, compared to the NB speech.

Table 1. *Mean transaction time in sec for each of the pictures. One standard deviation in parentheses.*

	NB	VOC	Babble	L2
Pic 1	266 (95)	371 (97)	338 (122)	*443 (120)*
Pic 2	244 (73)	333 (87)	303 (87)	*462 (156)*
Pic 3	262 (83)	313 (93)	*330 (109)*	*455 (103)*
Mean	**257 (74)**	**339 (81)**	**321 (106)**	**453 (126)**

3.1. Global acoustic-phonetic measures

The speech data and accompanying transcriptions can be used to investigate the acoustic-phonetic characteristics of casual and clear speech using a number of different measures. So far a number of global measures have been extracted including: mean word duration, fundamental frequency averages and range, long-term average spectrum measures (e.g. mean energy between 1 and 3 kHz, spectral slope). Further measures such as pause duration, proportion of hesitations, and modulation depth of the intensity envelope could also easily be made. Using the existing phonemic transcriptions, F1 and F2 range of vowels in all content words have been calculated and further vowel measures could be calculated.

3.2. Fine-grained segmental measures

The set of keywords in the corpus which have been produced in casual and clear styles in spontaneous and read speech allows for fine-grained analysis of acoustic-phonetic characteristics of the phonemes /p,b,s,ʃ/.

4. Applications of the DiapixUK task

The twelve picture pairs in the DiapixUK task can be used to investigate a range of research questions. The task is suitable for studies of various linguistic phenomena that are particularly prevalent in conversation, e.g. phonetic convergence. It is also suitable for investigating clinical populations. A small-scale study has already used a subset of the picture pairs to investigate relative communication difficulty in different cochlear implant simulations. Due to the digital layering of objects in the pictures, the scenes can be easily adapted to suit the needs of a particular project, e.g. different languages, accents. For example, some of the objects were changed in a subset of pictures to elicit target vowels for a small-scale study of phonetic convergence across two different British English accents. The task is also suitable for different age groups. The existing pictures were found suitable in pilot tests with 10 year olds and the pictures could be made simpler to suit an even younger or elderly audience.

5. Conclusions

In this paper, an overview of LUCID, a large corpus of spontaneous and read speech in casual and clear speaking styles has been presented, along with the DiapixUK materials which can be used to collect spontaneous speech in interaction. Due to the rich nature of the speech data in the corpus and the flexibility of the DiapixUK task, it is hoped that both the corpus and the task will be useful to other researchers for further research. To facilitate access to the corpus and the task materials, both will be available as part of an online archive facility that is currently under development at Northwestern University.

6. Acknowledgements

Thanks to our collaborator Ann Bradlow and colleagues at Northwestern University, who designed the original Diapix task, and to Martin Cooke for his advice on the babble condition. This project is funded by the UK Economics and Social Research Council (RES-062-23-0681).

7. References

[1] Ferguson, S. and Kewley-Port, D., "Talker differences in clear and conversational speech: Acoustic characteristics of vowels", J. Speech-Language-Hearing Res., 50:1241-1255, 2008.

[2] Smiljanic, R. and Bradlow, A., "Production & perception of clear speech in Croation and English", JASA,118(3):1677-1688, 2005.

[3] Smiljanic, R. and Bradlow, A.,"Speaking and hearing clearly: Talker and listener factors in speaking style changes", Ling & Lang Compass, 3:236-264,2009

[4] Blaauw, E. "The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech", Sp Comm, 14:359-374, 1994

[5] Hazan, V. and Baker, R., Does reading clearly produce the same acoustic-phonetic modifications as spontaneous speech in a clear speaking style? Proc. of DiSS-LPSS 2010, Tokyo, Japan, 2010

[6] Anderson, A. *et al.*, "The HCRC Map Task Corpus", Language and Speech, 34:351-366, 1991.

[7] Crawford, M., Brown, G., Cooke, M. and Green, P., "Design, collection and analysis of a multi-simultaneous-speaker corpus", Proc. Instit. of Acoustics, 16(5):183-190, 1994.

[8] Cooke, M. and Lu, Y., "Spectral and temporal changes to speech produced in the presence of energetic and informational maskers", JASA, conditionally accepted.

[9] Van Engen, K., Baese-Berk, M., Baker, R., Choi, A., Kim, M. and Bradlow, A., "The Wildcat corpus of native- and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles", Language and Speech, in press.

[10] Forster, K. and Forster, J., "DMDX:A Windows display program with millisecond accuracy", Beh. Res. Methods, 35:116-124.