Proceedings of DiSS-LPSS Joint Workshop 2010,
25-26 September 2010, Tokyo, Japan, pp. 07-10.

Does reading clearly produce the same acoustic-phonetic modifications as spontaneous speech in a clear speaking style?

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

Department of Speech, Hearing and Phonetic Sciences, UCL, UK
v.hazan@ucl.ac.uk[1], rachel.baker@ucl.ac.uk[2]

Abstract

This paper describes an acoustic-phonetic comparison of casual and clear speech styles elicited in read and spontaneous speech. For the spontaneous speech, 20 pairs of English talkers were recorded doing a problem-solving picture task in good and degraded listening conditions. Each person also read sentences in casual and clear styles. The read clear speech was an exaggerated form of clear speech relative to the spontaneous clear speech: it had higher median F0 in both styles, a greater increase in F0 range and greater decrease in speaking rate between casual and clear styles, and trends towards greater vowel space expansion.

Index Terms: spontaneous speech, read speech, clear speech, interaction, acoustic-phonetic characteristics

1. Introduction

The control that talkers have over the acoustic-characteristics of their speech (e.g., [1]) can be used to clarify speech to meet the needs of listeners. In this study, we investigate to what degree 'clear' speech which is produced when interacting with a talker with communicative difficulties differs from read clear speech produced without communicative intent.

Studies of clear or hyper-articulated speech have investigated which acoustic-phonetic features are enhanced when talkers attempt to maximize the clarity of their speech (for a full review, see [2]). This information is useful not only for gaining a better understanding of between- and within-speaker variability in speech production, but also for clinical or speech technology applications, as acoustically-enhanced speech may aid perception in difficult listening conditions. The Hyper-Hypo theory of speech production [3] is a useful framework for interpreting studies of clear speech: it suggests that talkers use the control that they have over their speech production to maximize communication efficiency while at the same time minimizing speaker effort. Hypo-articulated speech is adequate when there is a significant degree of linguistic-contextual information present, while hyper-articulated speech is produced in response to listeners' increased difficulty in understanding speech, e.g. in adverse listening conditions. The production of clear speech is therefore integral to the communicative process between talkers.

Perhaps surprisingly, methodologies commonly-used in clear speech studies do not involve communicative intent: talkers are recorded while reading sentences or words 'normally', and 'as if to a deaf person' or similar instruction (e.g., [4]). Alternatively, participants have been instructed to read sentences while doing a distractor task in order to elicit a 'reduced' speaking style or prompted to re-read sentences more carefully to elicit hyper-articulated speech [5]. These approaches produce carefully-controlled read speech recorded in optimal conditions; the acoustic-phonetic characteristics of this type of clear speech have been shown to be relatively consistent across studies. These include reductions in speaking rate [6], higher energy in the mid frequency region of the spectrum [7], higher mean fundamental frequency (F0) [6] and wider F0 range [6], longer and more frequent pauses ([8],[9]) and more expanded vowel spaces [8]. Some studies have shown changes in consonant/vowel intensity ratios [6] and in the modulation of intensity envelopes [7].

To understand if these acoustic-phonetic characteristics are representative of those of clear speech spontaneously produced in response to listener difficulty, a technique to elicit spontaneous clear and casual speech in a communicative situation is needed. One such technique is the Diapix task [10], a 'spot the difference' task in which two talkers are each presented with a picture and have to collaborate to find ten differences between the two pictures. This method shares many similarities with the well-known Map Task [11] but with a more balanced contribution between the two talkers within the task. The Diapix task can be controlled to naturally elicit clear speech in one or both talkers (e.g. [10]). In our study, we used a variant of the original Diapix task (DiapixUK) to elicit clear speech by placing a 'communication barrier' in the task. The communication barrier is placed on one talker only, so the conditions simulate that of a hearing person who is communicating with a cochlear implant user. We investigate how the talker who is hearing normally clarifies his or her speech in order to facilitate communication with the 'impaired' talker.

Although a number of studies have compared spontaneous and read speech, to our knowledge, no previous study has compared, for the same set of talkers, the acoustic-phonetic characteristics of two types of clear speech: clear read speech elicited through instruction and spontaneous clear speech elicited in dialogs involving communicative intent. The corpus described below is a subset of the LUCID corpus, reported in more detail in [12].

2. Method

2.1. Participants

Forty native speakers of Southern British English (20 M, 20 F; 19 - 29 yrs old) took part in the study. Participants, who were university students or faculty, volunteered with a friend of the same gender so there were 20 'friend' pairs. Participants read 'accent-revealing' sentences before being accepted onto the study, to ensure that they were from the same accent group; they were screened for normal hearing thresholds and reported no speech or language disorders.

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. Twelve pairs of pictures were designed, which were balanced in difficulty, as checked in a pilot study. These were four variants of three different themes, i.e. 4 beach scenes, 4 farm scenes, and 4 street scenes. Further details of this task can be found in [12].

2.2.2. Sentence reading task

A set of 144 sentences were designed, that contained a set of 36 keywords that were likely to be elicited within the diapix task (e.g., because the differences to be found involved one of those keywords). The sentences were meaningful, and of a simple syntactic structure, e.g. 'the old lady ate the *peach*/the young children loved the *beach*'.

2.3. Recording

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 [13]. Beyerdynamic DT297PV headsets were used in all recording sessions. In the diapix sessions, the speech of each talker was saved on a separate audio channel to facilitate transcription and acoustic analysis. Each participant was involved in five recording sessions. The first two were dialogue recordings with a friend (diapix task), and the last two were done alone (sentence reading). The data collected in the third session is not reported in this paper.

2.3.1. Diapix tasks

In the diapix task, 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 are asked to collaborate to complete the task; they communicate via headsets to locate the 12 differences between the two pictures. 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 the recording rooms and stopped the recording either once all differences were found, or after at least 15 minutes had lapsed and participants could not locate the final differences.

The diapix task was done in two test conditions: 'no barrier' and vocoder (VOC). In the 'no barrier' condition, both participants heard each other normally; they completed a diapix familiarization task, followed by three diapix tasks (1 beach, 1 farm and 1 beach scene). Both were asked to contribute to finding the differences to encourage a balanced conversation. The order in which pictures were completed at each session was counterbalanced across conditions and participant pairs using a modified Latin square design.

In the VOC condition, one of the talkers (the 'impaired' talker) 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 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 hearing normally, while still allowing enough communication to do the task. As there is a significant learning effect with vocoded speech, each participant first completed a ten-minute 'vocoder-familiarization' task. Each pair completed six diapix tasks in total: three when speaker 1's speech was vocoded and three when speaker 2's speech was vocoded. The speaker hearing normally was encouraged to take the lead in the conversation as it was this person's speech that was the focus of the analysis.

2.3.2. Read speech

The sentence reading task was presented using DMDX software [13]. After completing a picture naming task, which is not reported here, participants completed the reading task by reading sentences that appeared on a computer screen in a pseudo-randomized order. In session 1, they were asked to read the sentences 'casually as if talking to a friend'. In session 2, they repeated the task but were instructed to speak 'clearly as if talking to someone who is hearing impaired'.

2.4. Sound file post-processing

The mean duration of actual speech (i.e., linguistic material excluding silent pauses) for each of the 40 participants per condition is as follows: 'no barrier' diapix: 2.6 mins; VOC diapix: 4 mins; read casual speech: 4.1 mins; read clear speech: 6.5 mins. For all files, the speech of participant hearing normally was orthographically transcribed using freeware transcription software (Wavescroller) to a set of transcription guidelines based on those used by van Engen et al [10]. The transcripts were automatically word-aligned to the sound files using NUAligner software, which created a Praat TextGrid. The word-level 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.

2.5. Acoustic analysis

For the diapix recordings, a number of acoustic-phonetic analyses, highlighted in previous studies of clear speech, were carried out on the single-channel recordings for each picture and condition for each of the participants hearing normally. These include measures of: pitch range and median, mean energy between 1 and 3 kHz in the long-term spectrum, mean word duration and vowel formant range. For each measure and condition, the values were averaged over the three pictures to obtain a value per speaker per condition. For the read speech, analyses were carried out on the speech for each participant per condition, i.e. casual or clear.

2.5.1. Fundamental frequency median and range

Measures of median fundamental frequency and interquartile range in semitones were obtained using a Praat script, with a time step of 150 values per second. 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 speakers.

2.5.2. Long-term average spectrum (LTAS) measures

Silent portions were removed, using the silence annotations within the TextGrid; then the LTAS was calculated using a 50 Hz bandwidth, and the values for the first 50 bins (covering a 0-5000 Hz bandwidth) were calculated. A 1-3 kHz mean energy value was calculated using the bin values between these two frequencies.

2.5.3. Speaking rate (word duration) measures

The duration of each of the orthographically-annotated segments was obtained using a Praat script. Each annotated region was then tagged: agreement (AGR), breath (BR), filler (FIL), 'garbage' (GA) hesitation (HES), laughter (LG), silence (SIL) and speech (SP). Mean word duration was calculated by dividing the total time taken for the SP regions by the total number of words (SP regions) produced.

2.5.4. F1 and F2 vowel ranges

A Praat script removed annotations for all except content words in each file. Then, an SFS programme [14] was used to obtain a phonemic transcription of content words and a phoneme-level alignment to the waveform. Formant estimates were obtained in SFS for each phoneme segment and median vowel formant values (in ERB units) calculated for all monophthongs per talker per condition. Even though errors are possible at several stages of this analysis, the number of vowels used in the analysis and use of median values helps ensure that stable estimates can be obtained; the vowels used for the vowel range calculations were typically the most numerous. The range values were based, for each speaker, on the difference between the lowest and highest median F1 value and median F2 values across the vowel range.

3. Results

3.1. Comparison of transaction difficulty across spontaneous speech conditions

A measure of transactional difficulty was used to check that our VOC condition was successful in making communication more difficult between the two talkers. The time taken to find the first eight differences in the pictures was calculated (See Table 1). Task completion time was found to differentiate across talker groups in Van Engen et al (in press). The transaction time was significantly longer for the VOC condition than the 'no barrier' condition [$F(1,36)=64.4$; $p<0.001$], showing that longer interactions were required to complete the task in the VOC condition.

Table 1. *Mean time in seconds taken for the talker pairs to find the first eight differences for each of the pictures. Standard deviations are given in parentheses.*

	'No barrier'	VOC
Pic 1	266 (95)	371 (97)
Pic 2	244 (73)	333 (87)
Pic 3	262 (83)	313 (93)
Mean	**257 (74)**	**339 (81)**

3.2. Acoustic-phonetic measures

For each of the acoustic-phonetic measures examined, repeated-measures ANOVAs were run with task type (diapix, read) and speech style (casual, clear) as within-subject factors, and gender as between-subject factor, on the data obtained per talker, averaged across the three pictures per condition. Mean values across the 40 talkers for each measure in each condition are given in Table 2.

For median F0, the main effects of task type [$F(1,38)=24.1$; $p<0.001$] and speaking style [$F(1,38)=39.6$; $p<0.001$] were significant, with no significant interactions: median F0 was higher in read speech than in the diapix speech, and it was higher in the clear than in the casual speech. The between-subject effect of gender was significant.

For F0 range, the picture was more complex. There was a significant task type by speaking style interaction [$F(1,38)=6.4$; $p<0.05$]: there was a much greater increase in F0 range between the casual and clear styles in the read speech than the diapix speech. There was a significant task type by gender interaction [$F(1,38)=7.9$; $p<0.01$]: men showed a greater increase in F0 range than women in the read task relative to the diapix task. Further, there was a style by gender interaction [$F(1,38)=4.2$; $p<0.05$], with a greater increase in F0 range in clear speech in men than in women. Therefore, it seems that talkers have a higher pitch when reading than in conversation, and that, in both types of speech, they produce speech with higher fundamental frequency when speaking clearly. The increase in pitch range was more marked in the read clear speech than the clear speech produced in interaction, especially for male speakers.

For the mean energy 1-3 kHz measure, there was a significant task type by gender interaction [$F(1,38)=12.5$; $p<0.005$]: women had more energy in their speech for the diapix task than for the read task while men did not vary across task types. There was a main effect of speech style [$F(1,38)=17.4$; $p<0.001$] with higher mid frequency energy in the clear speech than the casual speech. No other effects or interactions were significant.

For word duration, the data analysis was carried out on the log-transformed data due to unequal variances in the raw data (See Figure 1). There was a significant task type by speech style interaction [$F(1,38)=43.6$; $p<0.001$]: mean word duration did not differ between diapix and read speech in the casual condition but was longer in the clear read speech than in the diapix VOC condition. Talkers therefore slowed down their speech to a greater extent when reading clearly than when clarifying their speech in interaction with another talker. The effect of talker gender was not significant.

In terms of vowel F1 range, there was a significant style by type interaction [$F(1,37)=6,14$; $p<0.05$]: there was a greater difference in F1 range between the casual and clear conditions for the read speech than for the diapix speech.

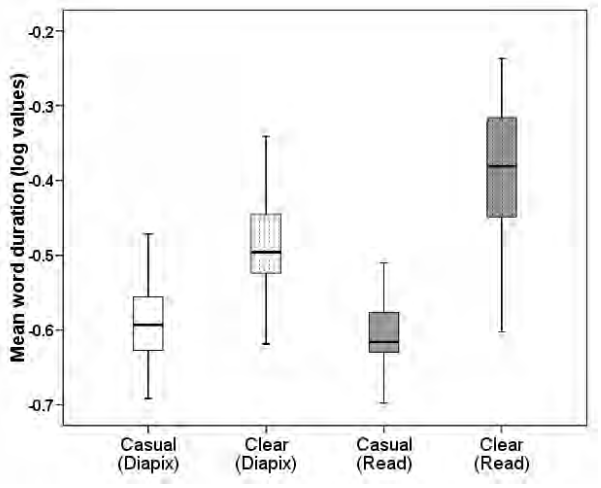

Figure 1: *Mean word duration (log values) for the diapix tasks ('no barrier' (casual) and VOC (clear) diapix conditions) and for the casual and clear read speech.*

	Diapix_'No barrier'		Diapix_VOC		Read_Casual		Read_clear	
	female	male	female	male	female	male	female	male
F0 median (sts)	91.5 (1.1)	80.4 (1.7)	92.2 (1.2)	81.6 (2.0)	92.3 (1.4)	81.0 (1.7)	93.1 (1.4)	82.7 (2.8)
F0 range (sts)	3.3 (0.8)	2.8 (0.6)	3.2 (0.7)	3.1 (0.6)	3.1 (0.8)	3.4 (0.9)	3.6 (0.9)	4.1 (1.0)
ME 1-3 kHz (dB)	25.6 (2.8)	23.6 (2.0)	27.5 (2.8)	25.4 (2.0)	22.9 (4.1)	23.4 (2.7)	23.4 (6.1)	25.7 (2.1)
Mean wd dur(ms)	266 (29)	250 (29)	345 (45)	310 (47)	252 (21)	247 (27)	413 (60)	428 (103)
F1 range (ERB)	3.5 (1.0)	3.8 (0.7)	4.7 (1.1)	4.4 (0.66)	3.8 (1.1)	4.0 (0.6)	5.0 (1.1)	5.3 (1.0)
F2 range (ERB)	5.0 (0.8)	3.5 (0.6)	6.1 (0.9)	5.3 (0.81)	6.0 (0.80)	5.7 (0.7)	7.9 (0.8)	8.1 (1.0)

Table 2. *Mean values (N=40) for a range of measures for speech produced in the diapix tasks ('no barrier' and VOC conditions) and in the read speech (casual, clear). Standard deviations are given in parentheses. The acoustic-phonetic measures are: median fundamental frequency and range (in semitones), mean energy in the 1-3 kHz region of the frequency spectrum, mean word duration, vowel F1 and F2 range (in ERB).*

There was a significant three-way interaction between task type, style and gender [$F(1,37)=5.9$; $p<0.05$]. For F2 range, the situation is more complex as there were significant interactions of task type with gender [$F(1,37)=21.9$; $p<0.001$]: men and women have a similar high F2 range for the read speech, but women had a larger F2 range for the diapix speech than men. There was also a significant speaking style by gender interaction [$F(1,37)=11.0$; $p<0.005$]: both men and women produced similar F2 ranges in the clear conditions but men produced a more reduced range in the casual condition than did women. Finally, there was a significant task type by speaking style interaction [$F(1,37)=10.7$; $p<0.005$], with a bigger difference in F2 range between the casual and clear read speech conditions than for no-barrier and VOC diapix conditions. Therefore, there was a tendency for the effect of speaking style on vowel range to be greater in read speech than in diapix speech, and women tended to use a more expanded vowel range in casual speech than men.

4. Discussion

In summary, there is evidence that clear speaking styles vary with task type. Read speech was produced with higher F0 median than speech produced in interaction in both speaking style conditions. Read speech also showed a greater change in F0 range between the casual and clear conditions than did the diapix speech, and this picture was more prevalent in men than in women's speech. The decrease in speaking rates across the casual and clear conditions was also greater in read than in diapix speech. It is only for the measure of mid-frequency energy that there was evidence of greater energy in the diapix than read condition (and this, for women only). Overall, therefore, the type of clear speech that is elicited when participants are asked to read sentence materials appear to be more extreme than the clear speech adjustments that talkers make in more naturalistic interactions produced in adverse listening conditions. This finding has implications for the interpretations of previous studies of clear speech especially as some (e.g., [8]) specifically selected speakers who had some voice training and were therefore likely to be at the 'clear' end of the range expected in a normal population even for casual speaking styles [15].

5. Conclusions

Clear speech produced in spontaneous speech with communicative intent shows similar acoustic-phonetic enhancements to those seen in clear read speech, but the degree of enhancement shown tends to be significantly higher in read speech for at least some measures. When considering what enhancements would be likely to be most useful in applications such as voice enhancers for clients with voice disorders or synthetic speech to be used in telecommunication systems, for example, basing these enhancements on studies of 'spontaneous' clear speech produced with situations in which the speech has communicative intent is more likely to produce clear speech which is natural-sounding and effective.

6. Acknowledgements

Thanks to our collaborator Ann Bradlow and colleagues at Northwestern University Linguistics Department. This project is funded by the UK Economics and Social Research Council (RES-062-23-0681).

7. References

[1] Johnson, K. and Mullennix, J., "Talker variability in speech processing", Academic Press, 1997.

[2] Smiljanic, R. and Bradlow, A., "Speaking and hearing clearly", Ling. Lang. Compass, (3):236-264, 2009.

[3] Lindblom, B. "Explaining phonetic variation: a sketch of H&H theory, in J. Hardcastle & A. Marchal [Eds], Sp. Production & Sp. Modelling, Kluwer Academic, 1990.

[4] Kain, A. et al., "Hybridizing conversational and clear speech to determine the degree of contribution of acoustic features to intelligibility", JASA, 124:2308-2319, 2008.

[5] Harnsberger, J. et al., "A new method for eliciting three speaking styles in the laboratory", Sp.Comm, 50:323-336, 2008.

[6] Bradlow, A. et al., "Speaking clearly for children with learning disabilities", J. Sp. Lang. Hear. Res. 46:80-97, 2003.

[7] Krause, J., and Braida, L., "Acoustic properties of naturally produced clear speech at normal speaking rates", JASA, 115:362-278, 2004.

[8] Picheny, M. et al., "Speaking clearly for the hard of hearing II", J. Sp. Hear. Res. 29:434-446, 1986.

[9] Liu, S., and Zeng, F., "Temporal properties in clear speech perception", JASA, 120: 424-432, 2006.

[10] Van Engen, K., et al., "The Wildcat corpus of native- and foreign-accented English", Lang. & Sp., in press.

[11] Anderson, A. et al., "The HCRC Map Task Corpus", Lang. & Sp., 34:351-366, 1991.

[12] Baker, R., and Hazan, V., "LUCID: a corpus of spontaneous and read clear speech in British English", this volume.

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

[14] Huckvale, M. et al., "The SPAR Speech Filing System", Euro. Conf. on Sp. Tech., Edinburgh, 1987.

[15] Hazan, V., and Markham, D., "Acoustic-phonetic correlates of talker intelligibility for adults and children", JASA, 116:3108-3118, 2004.