CLEAR SPEECH STRATEGIES AND SPEECH PERCEPTION IN ADVERSE LISTENING CONDITIONS

Jeremy Grynpas, Rachel Baker & Valerie Hazan

Department of Speech, Hearing and Phonetic Sciences, University College London (UCL), UK
jgrynpas@gmail.com; rachelbaker81@gmail.com; v.hazan@ucl.ac.uk

ABSTRACT

The study investigated the impact of different types of clear speech on speech perception in an adverse listening condition. Tokens were extracted from spontaneous speech dialogues in which participants completed a problem-solving task in good listening conditions or while experiencing a one-sided 'communication barrier': a real-time vocoder or multibabble noise. These two adverse conditions induced the 'unimpaired' participant to produce clear speech. When tokens from these three conditions were presented in multibabble noise, listeners were quicker at processing clear tokens produced to counter the effects of multibabble noise than clear tokens produced to counteract the vocoder, or tokens produced in good communicative conditions. A clarity rating experiment using the same tokens presented in quiet showed that listeners do not distinguish between different types of clear speech. Together, these results suggest that clear speaking styles produced in different communicative conditions have acoustic-phonetic characteristics adapted to the needs of the listener, even though they may be perceived as being of similar clarity.

Keywords: clear speech, speech perception, conversational speech, multitalker babble

1. INTRODUCTION

Speakers can attune their speech in adverse listening conditions for the presumed benefit of the listener. That is, in simulations of being in a noisy room, or in simulations of talking to someone with a hearing impairment, speakers are able to modify their phonetic output to clarify their speech [3]. A curious aspect of clear speech is that while there are a number of phonetic characteristics that distinguish conversational speech from clear speech (e.g., a decrease in speaking rate, wider dynamic pitch range, larger vowel space [5]), not all of these characteristics are ubiquitously employed. What is more, it is not clear that a single given acoustic cue used when producing clear speech has a beneficial effect on intelligibility for all listeners [5]. For instance, Burnham et al. [2] found that Australian English speakers exploited their vowel hyper-articulation, pitch and affect differently when speaking to babies, pets and other adults.

Hazan and Baker investigated clear speech strategies used in real communicative interactions between two speakers rather than elicited via instructions, as has been the case in many studies of clear speech [3]. More specifically, they investigated the possibility that the acoustic-phonetic characteristics of clear speech depend on the particular communication barrier that the interlocutor is trying to overcome. The acoustic-phonetic characteristics of the speech produced did indeed vary depending on the specific adverse listening condition. That is, speakers engaged in a problem-solving task with someone who was hearing them via a simulated cochlear implant (that only minimally transmits pitch information) made no changes to their F0 median and range (relative to their casual speech), presumably because enhancements in F0 characteristics were of little benefit in a condition where minimal F0 information was transmitted to their interlocutor. However, the same speakers did enhance these F0 features when interacting with someone hearing them in background noise [3]. Other clear speech features such as a slower speaking rate and increased F2 vowel range were common to both types of clear speech. These findings are in agreement with Lindblom's H&H model [4], which states that speakers modify their speech along a hypo-hyperarticulated continuum depending on the needs of the listener. Crucially, it appears that the speaker does not need to be directly exposed to the same adverse listening condition as the hearer in order to modify his or her clear speech strategy successfully.

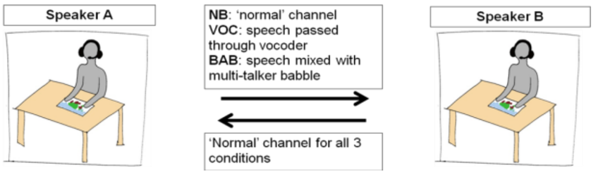
If it is indeed the case that a clear speech type is matched to a specific adverse listening condition (e.g. babble noise), then it should follow that 'clear' words produced to counteract babble noise should be more easily perceived when directly mixed with babble noise rather than 'clear' words produced to counteract another adverse listening condition (e.g. a simulated cochlear implant). In order to test this prediction, we presented listeners with tokens of conversational and two types of clear speech either produced to counteract the effect of multitalker babble (the simulation of a noisy room), or that of a vocoder (a rough simulation of talking to someone with a cochlear implant), all mixed with the same noise as was used in the multitalker babble condition.

In order to assess the subjective degree of clarity of these two different types of clear speech, listeners were also asked to rate different instances of conversational and the two above-mentioned forms of clear speech independent of adverse listening context. Our prediction was that if speakers are tailoring their phonetic output depending on the kind of adverse listening barrier they are trying to overcome, then listeners should be able to best perceive instances of clear speech when faced with the same adverse listening condition. Secondly, while we believe that listeners should be better able to identify the corresponding clear speech in this 'matched' condition, we do not expect that listeners will make a distinction in clarity ratings between the two types of clear speech in normal listening conditions. This can be best explained by H & H theory, under which it is reasonable to assume that a particular phonetic output is designed to counteract a given adverse listening environment, and so would be most suited to that (and only that) environment; H & H does not entail that one clear speech strategy is inherently more clear than the next.

2. METHOD

The stimuli were taken from the spontaneous speech recordings in the LUCID corpus, which involve pairs of speakers solving a 'spot the difference' picture task (diapix task) in different condition (Figure 1) [1]. Two speakers converse to find a number of differences between two variants of the same cartoon picture without seeing each other's picture. Speakers completed the task either when hearing each other normally (NB 'no barrier' condition), when one person's voice was passed through a three-channel noise-excited vocoder (VOC) or was mixed with eight-person multi-talker babble noise (BAB). The VOC and BAB conditions produced a communication barrier such that the person whose voice was degraded had to produce clear speech to help their interlocutor even though s/he was not directly experiencing the degradation. The three conditions provided spontaneous speech in casual (NB) and clear (VOC, BAB) speaking styles for 20 native southern British English speakers. Measures of communication efficiency (task transaction time, number of words produced) suggested that the VOC and BAB conditions did not differ significantly in terms of task difficulty [5]; that is, any differences in the clear speech produced was unlikely to be related to differences in task difficulty across the VOC and BAB conditions.

Figure 1: Figure showing how the 'adverse listening condition' affected only one of the two speakers (speaker B) thus forcing Speaker A to clarify his/her speech in order to successfully complete the problem-solving task.


Speaker A
NB: 'normal' channel
VOC: speech passed through vocoder
BAB: speech mixed with multi-talker babble
'Normal' channel for all 3 conditions
Speaker B

The differences in the diapix task are designed to elicit a set of keywords. In the LUCID corpus (using the DiapixUK picture materials), the differences across pictures relate to monosyllabic keywords beginning with either /b, p/ or /s, ʃ/. There are 9 /b-p/ minimal pairs, e.g. *bee*/*pea*, and 9 /s-ʃ/ pairs, e.g. *seat*/*sheet*. The 36 keywords are evenly distributed over the set of DiapixUK pictures [1].

A selection of keywords was extracted from the spontaneous speech recorded in the NB, VOC and BAB conditions (in VOC and BAB, tokens were used from the speaker who clarified their speech for their 'impaired' interlocutor). All instances of minimal pairs that were uttered in all three conditions at least once by a speaker were selected. Tokens which were whispered or coarticulated with a segment in an adjacent word were discarded. This yielded 53 minimal pair sets (26 /p/-/b/, 27 /s/-/ʃ/); a set contains one token of each word in the minimal pair from each of the NB, VOC and BAB conditions, i.e. 6 tokens in total. The tokens within each set came from the same speaker, but across minimal pair sets, tokens were produced by different speakers because not every speaker uttered each minimal pair in the three conditions. If there was more than one token suitable for inclusion in a minimal pair set, a random token was selected. In total, 156 /b/-/p/ and 162 /s/-/ʃ/ minimal pairs were used.

All speech files were normalized to a fixed intensity level then mixed with 8-talker babble noise at a signal-to-noise ratio of 0 dB, using a matlab script. This was the same noise as was used in the original recordings, and the SNR level was also close to that used in the diapix recordings for the LUCID database.

Thirty-seven native speakers of British English (27 female, 10 male, 18-30 yrs old) acted as participants in the perception study. They were monolingual, right-handed and reported no speech or hearing impairments. They were screened for hearing thresholds within 20 dB HL between 250 and 8000 Hz. All but five had normal hearing thresholds; the remaining five had slightly elevated thresholds in one or two frequencies.

The participants took part in two separate tasks, with each task separated into two blocks: one for /b/-/p/ tokens and one for /s/-/ʃ/ tokens. In the first task, the participants were presented with tokens mixed with babble noise, randomized across speakers and words. The participants were instructed to pay attention to the initial segment of the token and were told to press one of two keys on a computer keyboard corresponding to the initial consonant as quickly and accurately as possible but only after the whole word had been produced. A practice round consisting of 20 keywords from the diapix task (not included in the main stimuli set) that had been mixed in babble noise was undertaken to familiarize the participants with the procedure. Response keys were counterbalanced to minimize any handedness effects. The participants then participated in a ratings experiment. They were asked to judge the same word tokens presented in their original form, i.e. not mixed with noise, in terms of their clarity of production on a scale of 1-7, with ‘1’ indicating ‘very clear’ and ‘7’ indicating ‘unclear’. The whole experiment lasted 20-30 minutes.

For the identification task, the percentage of correct responses and mean reaction time (RT) for correct answers was calculated per speaker for /s/-/ʃ/ and /p/-/b/ words. RT was calculated from word offset to keypress. RTs above two standard deviations of the mean were excluded. For the ratings task, median ratings per speaker were calculated for each participant.

3. RESULTS

A high rate of correct responses was obtained given the relatively favourable signal-to-noise ratio that had been chosen to match conditions used in the LUCID database recordings. The rate of correct responses overall per condition for /s/-/ʃ/ was 91% (NB), 96% (VOC) and 95% (BAB); for /b/-/p/ tokens: 84% (NB), 84% (VOC) and 89% (BAB). A repeated-measures ANOVA showed a significant effect of contrast [F(1,35) = 46.1; p<.0001], with higher scores obtained for the /s/-/ʃ/ contrast, and a significant effect of condition [F(2,70) = 11.0; p<.0001], with a higher scores obtained for the BAB than NB tokens but no difference between BAB and VOC.

Given the ceiling effects obtained for the intelligibility score, reaction time provides a more sensitive measure of processing ease. Mean response times (RT) for both /b/-/p/ and /s/-/ʃ/ tokens were calculated for each of the three conditions. Generally, faster RTs were obtained for the **/s/-/ʃ/** contrast than for the /b/-/p/ contrast. The effect of condition on RT was significant for both the **/s/-/ʃ/** [F(2,72) = 131.1 p<.001], and /b/-/p/ [F(2, 70) = 54.0 p<.001] contrasts. For both contrasts, post-hoc tests showed that mean RTs were slowest for NB words, then VOC words, and fastest for BAB words (Figure 2), with each condition differing significantly from the others.

A separate task asked participants to rate the same tokens for clarity. The median scores for both **/s/-/ʃ/** and /b/-/p/ words showed a significant effect of condition (for **/s/-/ʃ/: [**F(2, 72) = 41.4 p<.001], for /b/-/p/: [F(2, 72) = 56.4, p<.001]. Clarity ratings were significantly higher (p<0.001) for words produced in the VOC/BAB conditions than for words produced in the NB condition (Figure 3). Unlike the response times, there was no significant difference in ratings between the VOC and BAB conditions for either **/s/-/ʃ/** or /b/-/p/ words.

4. DISCUSSION

In their acoustic-phonetic analysis of clear speech produced to counteract different adverse listening conditions, Hazan and Baker had suggested that the characteristics of the clear speech produced by a speaker varied according to the type of communication barrier imposed on their interlocutor, even if they themselves were not experiencing the adverse listening condition. A strong test of this hypothesis was to show that words produced in spontaneous speech aimed at counteracting the effects of multibabble noise were more easily processed by a listener when mixed with multibabble noise than when another type of clear speech (VOC) or casual speech (NB) were mixed with the same noise. Our RT data does indeed show this to be the case both for words with initial **/s/-/ʃ/** and /b/-/p/. This confirms our hypothesis that in speech communication, speakers are able to carefully attune their speech production to meet the needs of their interlocutor, as suggested by Lindblom’s H&H model [4] even if they themselves are not being exposed to the same adverse listening condition, and that these tailored adjustments benefit listeners more than other types of clear speech.

Figure 2: Bar chart showing the mean reaction times for correct response for the /b/-/p/ tokens (dark bars) and /s/-/ʃ/ tokens (light bars) for the three conditions (NB, VOC, BAB). The error bars show 95% confidence intervals.

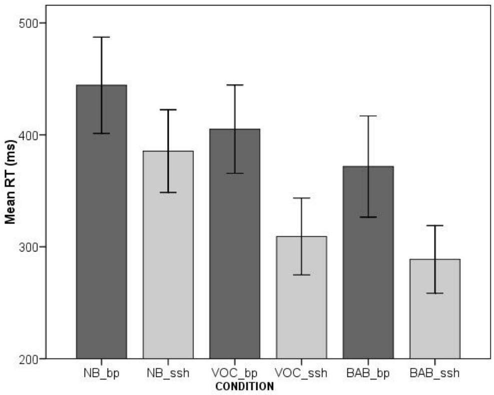
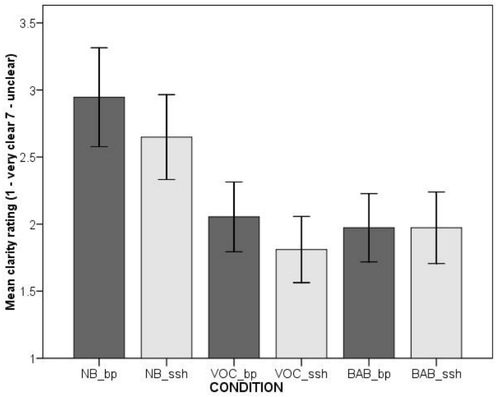


Figure 3: Bar chart showing the mean clarity ratings for the /b/-/p/ tokens (dark bars) and /s/-/ʃ/ tokens (light bars) for the three conditions (NB, VOC, BAB). The error bars show 95% confidence intervals.



The mean rate of correct initial consonant identification showed that higher responses were obtained for the ‘matched’ condition (BAB) than for tokens produced in the NB condition, but no significant difference between the BAB and VOC conditions. A ceiling effect was evident: that is, the level of babble noise was not high enough to show differences in consonant identification across the clear speech conditions, although the effect of condition was significant in terms of the more sensitive measure of reaction time. It would be interesting to investigate whether the effect of condition also influenced consonant intelligibility rates at higher noise levels.

The ratings study showed that both VOC and BAB spontaneous speech tokens were rated as clearer than the conversational speech tokens but that listeners rated both VOC and BAB tokens as equally clear. This shows that the faster reaction time obtained for words produced in the BAB condition are unlikely to be due to the fact that those tokens were just more hyper-articulated than those that had been produced in the VOC condition. It therefore seems to be the case that the VOC and BAB tokens are similarly clear, but have somewhat different acoustic phonetic characteristics adapted to the needs of the listener.

5. ACKNOWLEDGEMENTS

This project was funded by the UK Economics and Social Research Council (RES-062-23-0681).

6. REFERENCES

[1] Baker, R, Hazan, V. In press. DiapixUK: a task for the elicitation of spontaneous speech dialogs. *Behavior Research Methods*.

[2] Burnham, D., Kitamura, C., Vollmer-Conna, U. 2002. What’s new, Pussycat? On talking to Babies and Animals. *Science* 296, 1435.

[3] Hazan, V., Baker, R.S. Conditionally accepted. Acoustic-phonetic characteristics of clear speech produced with and without communicative intent. *J. Acoust. Soc. Am.*

[4] Lindblom, B. Explaining phonetic variation: a sketch of the H&H Theory. In Hardcastle, W, Marchal, A. (eds), *Speech Production and Speech Modelling.* Dordrecht, the Netherlands: Kluwer Academic, 403-440.

[5] Smiljanic, R., Bradlow, A. 2008. Speaking and hearing clearly: talker and listener factors in speaking style changes. *Language and Linguistics Compass* 3(1), 236-264.