Computer- and Human-Directed Speech Before and After Correction

Denis Burnham, Sebastian Joeffry, Lauren Rice

Marcs Auditory Laboratories, University of Western Sydney, Sydney, Australia

d.burnham@uws.edu.au, 10266058@student.uws.edu.au, l.rice@uws.edu.au

Abstract

Speech register research shows that humans are adept at fine-tuning components of their speech to accommodate the needs audience of the audience, suggesting that they have a model of human communication needs. However, when that audience is a computer rather than another human, such a model may be invalid. Here we examine humans' speech to other humans or an auditory-visual avatar before and after the computer or the human listener makes a listening "error". Speech is found to be hyperarticulated in Computer- compared with Human-Directed speech, and also in speech after correction. Results are discussed in terms of human-computer interaction and ASR systems.

Index Terms: computer-directed speech, speech repairs, vowel hyperarticulation, human-computer interaction.

1. Introduction

There are two distinct literatures concerning speech registers and speech adaptations for an audience: that concerning human-human interaction (HHI) and that concerning human-computer interaction (HCI) and each have particular methods of enquiry and dependent measures by which outcomes are evaluated. In this study we adopt elements of both. Relevant literature from each is provided below ahead of a synthesis and exposition of hypotheses.

1.1. Speech in Special Speech Registers

One of the most studied special speech registers is Infant-Directed Speech (Infant-DS), a style of speech that, compared to Adult-DS, has heightened fundamental frequency (F0) and wider F0 excursions, more repetitions, longer pauses and shorter utterances [1,2,3,4]. Research on Infant-DS has arisen from questions in language learning, e.g., to what extent does parental input facilitate infant and child language acquisition? Early studies were concerned with the verbal content of the speech but more recent acoustically- and phonetically-based analyses that have shown that Infant-DS has at least three distinct components, as follows:

- Attentional – measured by fundamental frequency (F0) level and variability [2]
- Affective – measured by ratings of speech that has been low-pass filtered to remove segmental and semantic content while maintaining prosodic information [5]
- Linguistic/Didactic – measured by area of a vowel triangle formed by joining the 1st and 2nd formant (F1, F2) values of [a], [i], and [u] vowels in F1/F2 vowel space [6,7].

Recently a range of special speech registers have been investigated with respect to these three components. Burnham [7] investigated both Infant-DS and Pet-DS (to dogs and cats) and found that both show heightened attentional and affective components compared to Adult-DS, but only Infant-DS showed vowel hyperarticulation, suggesting some didactic basis for Infant-DS. In a follow-up, Xu and colleagues [8] found evidence for vowel hyperarticulation in speech to parrots, supporting the notion that, if the audience is seen to be able to speak, then vowel hyperarticulation occurs.

In another series of studies the auditory input to the infant has been impoverished in order to simulate a hearing impairment. This is done in a double video set-up in which the mother is visible but her voice is modified either by lowering the amplitude [9,10] or simulating a hearing loss [11]. Under such simulated hearing loss conditions, mothers maintain the usual heightened pitch characteristics of Infant-DS but fail to hyperarticulate vowels. This suggests that for the didactic component of speech to be evident specific feedback from the infant is essential and that this feedback depends upon clear speech from the mother.

Finally, Uther et al. [12] have shown that, in contrast to the findings for Infant-DS and Pet-DS, Foreigner-DS is devoid of heightened pitch characteristics or heightened positive affect. Nevertheless, there is vowel hyperarticulation to foreigners compared with same-language adults, suggesting that vowel hyperarticulation is a didactic device *and* that the three speech components, attentional, affective and didactic, are separable and distinct. The results of these studies are summarised in the first 5 rows of Table 1.

Table 1: *Summary of special speech register studies for the attentional, affective, and didactic components of speech. The number of ticks (✓) and crosses (x) indicates the strength of each component for each speech register, and TBD stands for, 'To Be Determined'.*

Component / Audience	Attention (F0)	Affective (Ratings)	Didactic (V.HyperA)
Infants	✓✓	✓✓	✓✓
Infants (SimH'gLoss)	✓✓	✓✓	xxx
Pets (~Vocal)	✓✓	✓	xxx
Pets (Vocal)	✓✓	✓	✓
Foreigners	xx	xx (–ve)	✓✓
Computers	TBD	TBD	TBD

1.2. Speech in Human-Computer Interaction

Research on speech to computers has been driven by a different set of imperatives, most notably effective human↔computer interaction in practical or industrial user-directed applications. One solution to this problem is to design user interfaces so to constrain the human input to the computer and reduce possible errors at the ASR (Automatic Speech Recognition), natural language processing, and dialogue processing levels [13,14,15]. This is a worthwhile enterprise that promises improvements in effective HCI. However, another way of approaching speech in HCI is to aim for conditions in which the human is unconstrained, i.e., to make

ISBN 978-0-9581946-3-1
Accepted after peer review of full paper

HCI as natural and as similar to human-human interaction (HHI) as possible, i.e., as close as possible to Human-DS.

One of the essential components of such a naturalistic goal is to determine whether ASR systems can recognise human speech to computers, and what happens in response to errors – when the ASR system (i) fails to recognise (rejection), (ii) makes a false positive (insertion), or (iii) replaces input with an erroneous word (substitution). Such errors do not occur as frequently in Human-DS. When they occur in Computer-DS in HCI, the human usually makes repairs to try to have the computer understand them. A number of studies have investigated the types of repairs that are made under such conditions by manipulating the incidence of computer "errors" a human experiences when interacting with the computer [16,17,18,19]. For example, Oviatt, et al. (1998) [16] investigated speech modifications under both low and high computer error rates, and found that repairs included increased duration of utterances and segments, increased incidence and duration of pauses, decreased rate of speech (see also Stent et al. [17]), reduction of mean F0 in high computer error rate conditions, little change in F0 range (except for females in the high computer error rate condition), and no amplitude changes. Similar results have been found with children's repairs, with the addition of increased amplitude [18].

Phonological analysis has also revealed pre- to post-error repairs: Oviatt et al. found that human speech repairs involved more deliberate and well-specified speech including fewer disfluencies, and fewer reduced forms, e.g., 'fordy' → 'forty', 'tweny' → 'twenty', [16]; and Stent et al. found similar flap → released forms and reduced → full vowels in repairs [17].

The hyperarticulation in speech repairs is seen to fall along a continuum – Stent et al. [17] note that speakers return to their pre-error speaking style 4–7 utterances after misrecognition and both Stent et al. and Oviatt et al. note that there are similar but gradually-accumulating features of repairs between low and high computer error rates. Oviatt et al. [19] has formalised this in the Computer-elicited Hyperarticulate Adaptation Model (CHAM). In this two-stage model, Stage I repairs (at low error rates) involve solely limited increases in durational characteristics; then in Stage II (at high error rates) repairs involve more extended durational changes and changes in articulatory and F0 aspects of speech.

These studies have been conducted in a different tradition and using different dependent variables to those used in the Infant- and Pet- and Foreigner-DS studies. This is reflected in the last row of Table 1, in which entries for the attentional, affective and didactic components of this special speech register have been assigned a 'To Be Determined' value. The purpose of this study is to go some way towards filling in these cells.

In this study durational and F0 characteristics of repairs are investigated, but instead of measuring articulatory repairs by lack of reductions etc. [16,17], the Infant-, Pet, and Foreigner-DS vowel triangle area method is used to measure hyperarticulation. In previous HCI studies hyperarticulation is associated with Stage II (CHAM) functioning in which there is a high, 20% [16], 33% [17], rather than a low 6.5% [16], 8.3% [17] computer error rate (where the error rate refers to the percentage of sentences in which there was an "error". In this study, a computer error rate of 33% was used in order to optimise the possibility that vowel hyperarticulation would occur in speech repairs.

This study differs from previous HCI studies in two ways: first, in addition to the HCI condition, an HHI condition was included in which the user spoke to another human; and second in concert with Infant-DS-type studies but not Computer-DS-type studies, both the HCI and HHI presentations were auditory-visual – in HCI users spoke to an avatar, and in HHI conditions users spoke to another human with all other conditions kept constant.

It was expected that there should be higher mean F0, greater pitch range, longer vowel durations, and greater vowel hyperarticulation as a function of both manipulated factors, i.e., both (i) after than before repair and (ii) in Computer-DS than in Human-DS condition. There were no expectations about whether there would be any interaction between these two factors.

2. Method

2.1. Participants

Twenty-four Introductory Psychology University of Western Sydney undergraduates (12 women and 12 men Mean = 23.74 years) participated and received course credit.

2.2. Design

A 2x2 (speech register (Human-DS/Computer-DS) x Time (Before/After computer error) within-subjects design was used with 4 dependent variables, vowel triangle area, vowel duration, mean F0, and F0 range. Counterbalancing controlled for order and stimulus materials.

(a) **(b)**

Figure 1: *(a) Participant's view of the Thinking Head (or human in the other condition) display (b) Experimenter's control room set-up for the Wizard of Oz.*

2.3. Apparatus

A 'Wizard of Oz' paradigm was used employing Double Video equipment [9,10,11] and two test rooms. The participant sat in one, facing a monitor on which they could see and interact with a talking head avatar or a human (depending on the condition) both located in the second room (see Figure 1a) The experimenter drove the avatar using a menu-based system to present, via the avatar, oral questions, acknowledgements and computer "errors" (see Figure 1b). In the Human-DS condition, the participant saw the experimenter on their screen. The experimenter followed a script the same as that for the avatar except for the items (counterbalanced across participants). The avatar was the Head-0 Thinking Head, an interactive embodied conversational agent (ECA) that nods, blinks, and responds to human input based on a Chatbot engine via AV integrated speech using Festival Speech Synthesis with a male voice speaking in a British accent [20]. For further Thinking Head details [see 21, 22].

2.4. Procedure

The Stent et al. [17] procedure was adapted for use here. Prior to the trials starting, participants were provided with a list of names of protagonists with the suburb in which they lived, and items each would bring to a bake sale, and to a garage sale. There were then 3 question types based on these (from the avatar or the human depending on condition), in which participants were required to answer in a prescribed format, e.g.,

Q: What is Barry bringing to the garage sale?
A: Barry is bringing a 'dirty deeds' DVD to the garage sale.
Q: Did you say Barry is bringing a 'dirty deeds' DVD to the garage sale?
A: Yes, Barry is bringing a 'dirty deeds' DVD to the garage sale.

On Error-Repair trials the human or computer made substitution and rejection errors. Substitution errors (which account for over 90% of recognition errors [23]) were used to investigate vowel hyperarticulation, and included errors on the 3 corner vowels /i/, /u/, /a/, e.g.,

Q: Did you say Barry is bringing a 'dry deeds' DVD to the garage sale?
A: No, Barry is bringing a 'dirty deeds' DVD to the garage sale.

Rejection errors were included to examine global repairs and in these the participant was simply asked to repeat the answer.

There were 72 questions, 36 in each register, separated into 3 blocks of 12. In each block there were 3 Substitution-Repair trials (1 for each corner vowel, /i,u,a/, 3 for each vowel per condition), and 1 global Rejection-Repair trial. So the vowel-specific substitution computer error rate was 25% of the sentences presented and a global computer error rate, 8.3% of the sentences presented, an overall error rate of 33%. Global errors were included to provide some variation, and only data from the substitution error trials are analysed here as it is specifically vowel hyperarticulation that is of interest.

3. Results

Audio recordings of the target words from Audacity V1.3.7 were exported as a mono .wav files at 44Khz 16bits. The corner vowels in these target words were segmented manually via PRAAT V 5.1.12, and formants and vowel durations extracted. Files with undefined formant values due to mis-tracked formants were deleted, leaving 845 files for analysis across the 4 conditions. F1 and F2 values for the vowels in each condition were averaged allowing four triangles to be plotted in F1/F2 space. The data were converted to the perceptual Mels scale and vowel triangle areas for each participant in each of their four conditions were used for statistical analysis. Mels was used as the measure as it is the intended effect of the speech on the audience of importance here, so it is the *perceptual* rather than the acoustic aspects of the speech that are important to analyse. Moreover, this is the scale that has been used as the standard in pioneering studies concerned with Infant-DS [6]. For F0 mean and F0 range, the entire sentences featuring the target words were used to retain intact prosodic information, and mean and range F0 extracted via PRAAT. Pitch range was converted from absolute Hertz values to ratio pitch values using the semitone scale. When strictly applied, this transformation should be conducted on all measures of fundamental frequency. However, studies of Infant-DS tend to use semitones for range and Hertz for other measures [1]. Participants with insufficient data for phonetic analyses were excluded from the statistical analyses. On this basis, 4 sets of data were excluded from the vowel duration, and 4 sets from mean and range F0. The results were analysed using standard analysis of variance procedures. For each of the four dependent variables (vowel triangle area, vowel duration, mean F0 and range F0) 2 speech registers (HCI/HHI) x 2 conditions (before/after) analyses of variance (ANOVAs) with repeated measures on both factors were conducted.

3.1. Vowel Triangles

Vowel triangle areas are shown in Figure 2. ANOVA revealed that Computer-DS had larger vowel triangle areas than Human-DS, F(1,23) = 11.10, p = .003, partial η2 = .33, and speech after larger areas than that before correction, $F(1,23) = 4.70$, $p = 0.04$, partial η2 = .17; but no significant interaction, $F(1,23) = .02$, $p = 0.89$, partial η2 = .009.

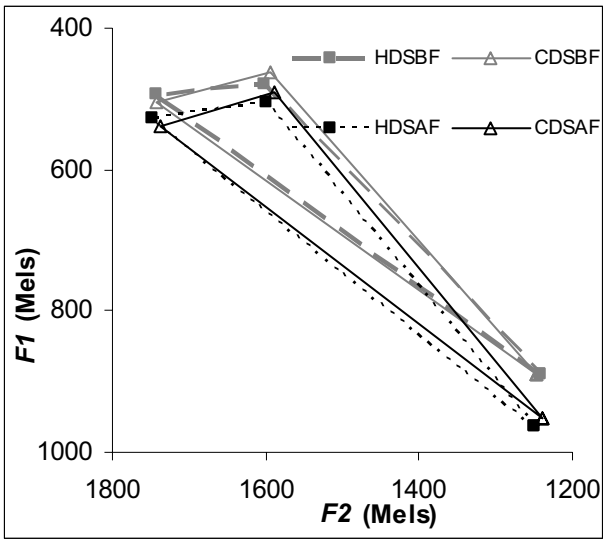

Figure 2: *Vowel triangle areas for speech to Humans (HDS) and Computers (CDS) before (BF) and after (AF) correction.*

3.2. Vowel Duration

Vowel durations are shown in Table 2. Computer-DS speech had longer vowel durations than Human-DS, F (1, 20) = 5.62, $p = .03$, partial η2 = .24 and speech after correction longer than speech before correction, F (1, 20) = 15.824, $p > .0$, partial η2 = .01. There was no significant speech register x correction interaction F (1, 20) = .38, p = .55, partial η2= .05

Table 2: *Means(*standard errors*) for vowel duration, F0 mean & range of human- directed speech (HDS) & computer-directed speech (HDS)before & after correction.*

	Vowel Duration (secs)	F0 Mean (Hz)	F0 Range (semitones)
HDS Before	.151 *.008*	217.275 *12.582*	15.607 *1.040*
HDS After	.177 *.012*	219.693 *11.365*	17.174 *0.947*
CDS Before	.140 *.008*	216.960 *10.629*	15.929 *1.214*
CDS After	.161 *.013*	228.574 *12.002*	16.909 0.833

3.3. Mean and Range F0

Mean and range F0 values are shown in Table 2. For speech register there was no significant Computer-DS vs. Human-DS difference for either measure (F_{MeanF0} (1,20) = 0.19, $p = .67$, partial η2 = .0; ($F_{RangeF0}$ (1,20) = 2.05, $p = .17$, partial η2 = .09). For speech correction there was significantly greater range F0 after correction ($F_{RangeF0}$ (1,20) = 6.25, $p = .02$,

partial η2 = .24) but no differences for mean F0 (F_{MeanF0}(1,20) = 2.05, p = .17, partial η2 = .09). There were no significant interactions (F_{MeanF0}(1,20) = 1.17, p = .29, partial η2 = .05); ($F_{RangeF0}$(1,20) = .27, p = .61, partial η2 < .01).

4. Discussion

4.1. Speech Before and After Correction

Computer correction affected subsequent speech repairs. In accord with previous findings, after correction users' repairs showed hyperarticulated vowels, longer-duration vowels and an increase in range F0. However, in contrast to previous findings there were no effects of computer correction on mean F0 [16,17]. In Oviatt's model, CHAM [19], Stage I of hyperarticulation involves only durational elongations [16]. Then in Stage II, which is only activated under high computer error rate conditions, further durational elongation *plus* articulatory and F0 aspects of hyperarticulation are evident. The fact that after correction there was extended duration, increased F0 range (although not for mean F0) and vowel hyperarticulation suggest that the relatively high computer error rate here (33% overall and 25% for substitution errors, which are similar to the high rates in other studies – 20% [16] and 33% [17]), successfully took participants into this second stage. These results are encouraging, as they suggest comparability with other studies using the computer correction paradigm. They also show the generality of the effect of computer correction; in previous studies articulatory hyperarticulation after correction involved fewer disfluencies, and fewer reduced forms, whereas here articulatory hyperarticulation involved the measure most often used in the special speech register studies, extended vowel triangle areas.

4.2. Talking to Computers and Humans

The most interesting finding of this study was that there was vowel hyperarticulation across the board for Computer-DS compared to Human-DS. So a tick or two can be placed in the bottom right hand cell of Table 1 – vowel hyperarticulation for Computer-DS. This was accompanied by longer vowel duration in Computer- than Human-DS especially after correction, but no greater F0 (attentional) characteristics, so crosses can be placed in the bottom left hand cell of Table 1. (Studies of the affective aspects of Computer-DS are yet to be conducted.) These results show that participants appear to treat the avatar as an entity requiring special treatment and in this respect there are two possibilities: the avatar may be treated like an infant or a foreigner (see Table 1). As there was no elevation of F0 characteristics in Computer-DS, it appears that the latter is the case – rather than being addressed with elevated mean F0 and greater pitch modulation as would a "cute" baby, the avatar was rather treated as an adult who could not speak or hear properly (akin to a foreigner). Accordingly, longer duration and hyperarticulated vowels, but not elevated mean F0 or pitch range, were employed.

Ratings of affect were not included in this study but additional ratings studies and analyses are currently underway. The results of such studies will be of interest in filling in the final cell in Table 1. In addition, further studies in which the characteristics of the avatar (e.g., degree of smiling, interactivity, etc.) are varied would be useful, as would a condition in which the degree of the 'foreignness' of the avatar's speech was systematically varied in order to evaluate the relative contribution of Foreigner-DS and Computer-DS to the hyperticulation found here. Finally, a control condition in which there is auditory-only exposure to the avatar's (and the human's) voice would be of use in order to determine the locus and origin of modifications in Computer-DS.

5. Conclusions

The results reinforce previous findings by showing that hyperarticulation occurs in speech repairs after computer correction. This hyperarticulation involves vowel lengthening and increases in range (but not mean) F0, and now an additional variable, vowel hyperarticulation (as measured by extended vowel space), that has not been included in previous studies.

Most importantly, the results show for the first time, that when speaking with a somewhat lifelike avatar compared with speaking to another human adult, users lengthen and hyperarticulate their vowels just as they would when speaking to a foreigner. Whether such adjustments are useful for particular ASR systems is yet to be determined and so studies running the Computer-DS and the Human-DS speech through an ASR system such as Sphinx or Nuance would be informative.

The human-computer speech comparison method used here brings together special speech register and HCI research and should prove to be a useful paradigm for further fine-grained studies. In particular, investigation of avatar characteristics that promote hyperarticulated speech will be of great importance not only for HCI, but also for determining the critical features of the audience that promote hyperarticulation to infants, parrots, foreigners and other recipients.

6. Acknowledgements

1. This research was supported by an ARC/NH&MRC Thinking Systems grant to Burnham et al. (TS0669874).
2. We appreciate the input from Dr Christian Kroos in writing the Wizard of Oz software, and Mr Steve Fazio in setting up the Wizard of Oz and Double Video hardware.

7. References

[1] Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: Is the melody the message? Child Dev., 60, 1497-1510.

[2] Fernald, A., & Kuhl, P. (1987). Acoustic determinants of infant preference for motherese speech. Inf. Behav. Dev.,10, 279-293.

[3] Fernald, A., & Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. Dev.Psych., 20, 104-113.

[4] Grieser, D., & Kuhl, P. (1988). Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese. Dev. Psych., 24, 14-20.

[5] Kitamura, C., & Burnham, D. (2003). Pitch and communicative intent in mother's speech: Adjustments for age and sex in the first year. Infancy, 4, 85-110.

[6] Kuhl, P. K., Andruski, J. E., Chistovich, I. A., et al., (1997) Cross-language analysis of phonetic units in language addressed to infants. Science, 277, 684-686.

[7] Burnham, D., Kitamura, C., & Vollmer-Conna, U. (2002) What's New Pussycat? On talking to babies and animals. Science, 296, 1435.

[8] Xu, N., Burnham, D., Kitamura, C. & Vollmer-Conna, U. Polly Want a Cracker? On talking to babies, puppies and parrots. (2004) Poster, International Conf. Infant Studies, Chicago, USA.

[9] Lam, C., & Kitamura C. (In Press). Maternal Interactions with a Hearing and Hearing-impaired Twin: Similarities and

Differences in Speech Input, Interaction Quality, and Word Production, J. Speech Language & Hearing Research.
[10] Lam, C., & Kitamura C. (2009). Infant-directed speech to infants with a simulated hearing loss, JASA, 125, 2533.
[11] Rice, L. & Burnham, D. (2010) Speak clearly mummy! The effect of degraded input on infant-directed speech. Australasian Experimental Psych. Conf., April 8-10, Univ. Melb. Australia.
[12] Uther, M., Knoll, M.A., and Burnham, D. (2007) Do you speak E-N-G-L-I-SH? A comparison of foreigner- and infant-directed speech. Speech Communication, *49*(1), 2-7
[13] Hone, K., & Baber, C. (1999). Modelling the effects of constraint upon speech-based human–computer interaction. International J.Human-Computer Studies, 50, 85-107.
[14] Hone, K., & Baber, C. (2001). Designing habitable dialogues for speech-based interaction with computers. Internat. J. Human-Computer Studies, 54, 637-662.
[15] Oviatt, S., Cohen, P., Wu, L., Duncan, L., et al. (2000). Designing the user interface for multimodal speech and pen-based gesture applications: State-of-the-art systems and future research directions. Human-Computer Interaction, 15, 263-322.
[16] Oviatt, S., MacEachern, M., & Levow, G. (1998). Predicting hyperarticulate speech during human–computer error resolution. Speech Comm, 24, 87-110.
[17] Stent, A. J., Huffman, M. K., & Brennan, S. E. (2008). Adapting speaking after evidence of misrecognition: Local and global hyperarticulation. Speech Comm, 50, 163-178.
[18] Oviatt, S., and Coulson, R. (2003) Predicting children's hyperarticulate speech during human-computer error resolution. JASA, 113, 2296.
[19] Oviatt, S., Levow, G., Moreton, E., & MacEachern, M. (1998). Modeling global and focal hyperarticulation during human–computer error resolution. JASA, 104, 3080-3098.
[20] Black, A. & Taylor. P. (1997) The Festival Speech Synthesis System: System documentation. Tech. Report HCRC/TR-83, Human Communication Research Centre, Univ. of Edinburgh.
[21] http://thinkinghead.edu.au/
[22] Burnham, D., Abrahamyan, A., Cavedon, L., Davis, C., Hodgins, A., Kim, J., Kroos, C., Kuratate, T., Lewis, T., Luerssen, M., Paine, G., Powers, D., Riley, M., Stelarc, Stevens, K. (2008) From Talking to Thinking Heads: Report 2008. In Göcke, R., Lucey, P., and Lucey, S. (Eds) *Proceedings of the International Conference on Auditory-Visual Speech Processing 2008*, 127-130. Adelaide, Causal.
[23] Baber, C., & Hone, K. (1993). Modelling error recovery and repair in automatic speech recognition. Internat.J Man-Machine Studies, 39, 495-515