THE EFFECT OF ARTICULATORY REDUCTION ON INTELLIGIBILITY AT FAST SPEECH RATES: WORD RECOGNITION IN NATURAL FAST SPEECH VS. COMPRESSED SLOW SPEECH

Charles Chang
Linguistics 210
Prof. Keith Johnson
Final Paper

10 May 2005

1. Introduction

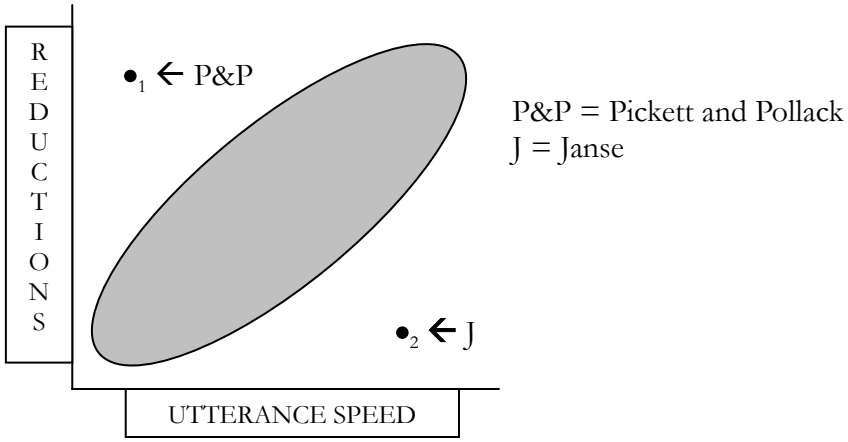
1.1. The Null Hypothesis

It might be a natural prediction that, *ceteris paribus*, the intelligibility of speech decreases as the rate of utterance or extent of articulatory reduction increases: a listener should have a harder time processing an utterance when a talker is speaking quickly or slurring their speech. This prediction is supported by one result of Pickett and Pollack (1963), who show that temporally stretched fast speech samples are much less intelligible than non-stretched slow speech samples of approximately the same length (163); in other words, at a constant utterance length and speed, the increased reductions of fast speech render it less intelligible.

1.2. Counterevidence

However, Pickett and Pollack (1963) also find that for a given utterance speed, a temporally stretched version of an utterance is consistently less intelligible than the non-stretched version; thus, it would appear that intelligibility can actually suffer at a slower speech rate. In addition, Janse (2003), working with synthetic and natural speech in Dutch, shows that intelligibility of hyperarticulated synthetic diphone speech suffers more than that of natural speech when both are time-compressed (63), suggesting that at a fast speech rate, speech with reductions is more intelligible than speech without reductions. Soltau and Waibel (1998) find that for speech recognizers as well, hyperarticulated speech can decrease recognition accuracy. It would thus appear that intelligibility can suffer in the presence of more robust articulation, also. These observations are depicted in the schematic below, with the first point marking the result of Pickett and Pollack and the second point marking that of Janse.

(1) Schematic of relation between utterance speed and speech reductions (idealized)



Taking into account natural variability among talkers in the amount of reduction they effect in conversational speech, there appears to be a "sweet spot" of intelligibility (the shaded oval in the figure above) that shows a somewhat linear relation between speech reductions and utterance speed: the faster a talker speaks, the more s/he is compelled to reduce articulation in the interest of maintaining intelligibility for the listener—an effect of communicative empathy (cf. Lindblom 1990). If a talker ventures outside this region—hypoarticulating at a slow speech rate or hyperarticulating at a fast speech rate—then intelligibility suffers.

1.3. Against Articulation as Interference

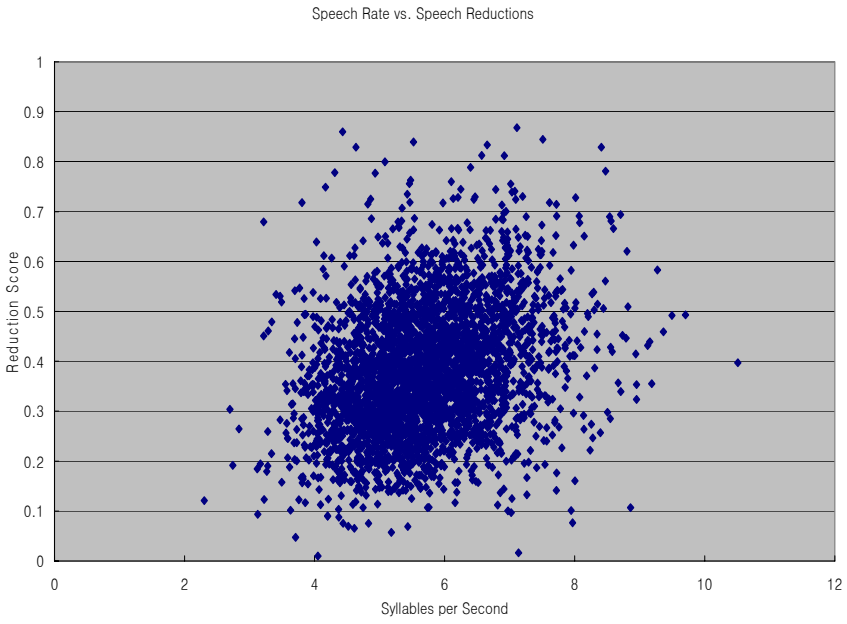
On the other hand, the results of Janse (2003, 2004) call into question the validity of the above conclusions. Janse finds, on the contrary, that word perception in artificially compressed normal speech is easier than in naturally produced fast speech. Thus, the second point in figure (1) may not lie outside the region of intelligibility, but rather inside it; the region of intelligibility, then, may not be limited to a linear oval, but may extend beyond the oval toward the horizontal axis. In other words, more robust articulation (i.e., fewer reductions) may increase intelligibility no matter what the speech rate.

1.4. Present Study

The question remains: is there really a linearly oriented region of intelligibility such that articulatory cues are helpful at slow speech rates but harmful at fast speech rates? Or is it instead the case that these cues are always helpful to listeners at any speech rate? Using speech samples from the Variation in Conversation (ViC) corpus of conversational speech, this study aims to investigate this question.

Data from the ViC corpus conform quite well to the estimates in figure (1).

(2) Plot of utterance speed (syl/sec) vs. speech reductions[1] in the ViC corpus



However, the above graph is a representation of talkers' behavior; it is unclear whether this

[1] The amount of speech reduction is represented by a score out of 1.0 taking into account the number and kind of reductions that take place over the course of an utterance.

corresponds directly to a region of intelligibility with respect to listeners. The present study will investigate the extent to which the rate/reduction correlation in conversational speech seen in figure (2) is the result of conformity to a region of intelligibility rather than of limits on human articulatory capacities. Are points missing in the lower right-hand portion of the graph (fast utterance, low reduction) because talkers choose to omit cues that may overload listeners' processing capacity and decrease intelligibility? Or do talkers not venture into this region because they simply cannot produce speech at such high speeds without reducing their articulation?

Janse's (2003) results with synthetic diphone speech suggest that the former explanation may account for this gap, while her (2003, 2004) results with natural speech suggest that the latter explanation is actually more accurate. This study undertakes a perception experiment with non-compressed natural fast speech and compressed natural slow speech to see if the latter explanation is really the right one.

2. Methods

2.1. Subjects

The talker subject was a male native speaker of English with no hearing or speech impediments. All listener subjects were also native speakers of English with no hearing or speech impediments. There were a total of 13 male and female listeners split across two groups, seven in Group A and six in Group B.

2.2. Materials

All recordings were made with a Sony Vaio PCG-TR5L laptop microphone, and all speech samples were played to listeners on a Sony Vaio PCG-TR5L laptop computer over Direct Sound EX-29 noise reduction headphones. All time compression of speech samples was done in Praat.

2.3. Procedure

2.3.1. Preparation of Materials

Speech samples of several types were first extracted from the corpus[2]: fast reduced utterances, fast unreduced utterances, and slow unreduced utterances ranging from 10 to 17 words in length. Seven particularly fast, particularly reduced utterances (as well as slow and fast filler utterances and a control utterance) were selected for inclusion in the perception experiment, and a talker was recruited to record the seven utterances at a slow speech rate. He was instructed to say the utterances as if he were speaking to a beginning L2 learner of English.

The length in seconds of each of the seven extracted utterances, as well as the corresponding recorded utterance, was noted in Praat. The ratio between the length of the fast extracted utterance and the slow recorded utterance was first calculated, and this ratio was input as the factor in Praat's linear time compression feature. In this way, a time-compressed version of a slow recorded utterance was created that matched the corresponding fast extracted utterance in length.

Each utterance was then gated into incrementally longer sections of words beginning with the first word of the utterance (cf. Pickett and Pollack 1963). The talker subject was used as an objective listener to judge the placement of the gate boundaries. In this way, each utterance was divided up into as many gates as it contained words, with the final gate containing the entire utterance.

Two perceptual tests were then devised which contained a mix of natural fast and slow

[2] Thanks to Keith Johnson for his help with this task.

utterances and time-compressed utterances. Both tests contained 12 different utterances. The test presented to Group A listeners contained four time-compressed utterances, three natural fast reduced utterances, two natural fast unreduced utterances, two natural slow unreduced utterances, and a control utterance. The test presented to Group B listeners contained three time-compressed utterances, four natural fast reduced utterances, two natural fast unreduced utterances, two natural slow unreduced utterances, and a control utterance.

2.3.2. Perception Experiment

Using the materials that were prepared as described above, a perception experiment was run with 13 native English listener subjects using an experimental design incorporating elements of Pickett and Pollack (1963) and Shockey (1998). Critically, listeners did not perform a timed phoneme identification task as in Janse (2003, 2004), but a word identification task instead. Subjects were divided into two groups to which a different set of 12 sentences were presented in gates of increasing length; thus, for a particular sentence subjects first heard a one-word gate (containing the first word), followed by a two-word gate (containing the first and second words), and so on until the entire sample was played.

After a particular gate was played, subjects wrote down the words that they had heard on an answer sheet containing a row of blank spaces for each word in that gate; after they finished writing down their responses, the following gate was played. Subjects were allowed as much time as they needed to write down their responses. Before beginning the test, they were told that the utterances they would hear might or might not be full sentences, and that they could change their mind about what they heard in a previous gate on a following gate; however, they were instructed to leave their responses to previous gates in tact and to write a new guess about a particular word only on the row for the current gate. In this way, the pattern of change in subjects' perception of a sentence as gate length was increased was recorded on their answer sheets. Each trial lasted approximately 20 minutes, and subjects were compensated for their efforts with various and sundry sweets.

2.3.3. Quantification of Responses

Subjects' responses were scored on a scale of 10 as follows. Successful perception of the target word, as indicated by a matching written response on the answer sheet, at the gate in which the word was introduced (i.e., the first gate that perception of the word was possible) was given a maximum score of 10. For every additional gate that was needed to perceive the word correctly, one point was subtracted from this score; thus, a subject that needed three additional gates to perceive a given word correctly was given a score of 7 for that word.

Misspelled responses, as long as they were homophonous with the target word, were given full credit, and responses that were not exactly right but contained the whole morpheme in question (e.g., the response *looking* for the target word *look*) were also given full credit. Otherwise, responses that failed to reproduce the target word before the last gate were scored as zero.

2.3.4. Statistical Analysis

Scores for a particular word across subjects, in addition to scores for one subject across all words in an utterance, were averaged and then subjected to statistical analysis. A double-sided p-value was calculated online (at http://home.clara.net/sisa/t-test.htm) for the performance of Group A subjects and Group B subjects on each pair of natural fast utterance and time-compressed slow utterance, with p-values less than 0.05 being taken as indicating a statistically significant result.

3. Data

Full tables of individual scores and averages are presented below. Note that 'S' stands for 'subject' and 'AVE' stands for 'average'; 'x' marks a time-compressed utterance. Each utterance is labeled with the speaker from the ViC corpus who produced it and the time in the corpus at which it began. The data table for each utterance is indexed in the uppermost left-hand cell by the subject group (A or B) to which the utterance was presented, as well as the temporal position (among 12 utterances) in which it was presented. As described above, higher average scores for a word, utterance, or subject indicate a greater degree of ease and fidelity in perception, while lower scores suggest degraded perception.

3.1. Control Utterance

Since listeners were expected to vary in their processing capabilities, one of the utterances on Group A's test and on Group B's test was kept identical in order to ensure that Group A and Group B were equally matched. The data for this control utterance are presented in tables (3) and (4) below.

(3) Group A responses to utterance of s3202b, 291.988 sec (control)
(rate = 9.18 syl/sec, reduction = 0.36)

A6	there's	no	reason	to	Be	scared	of	anything	that	you	do	AVE
S1	8	9	10	9	10	10	9	10	9	10	10	9.45
S2	10	10	10	10	10	10	10	10	10	10	10	10.00
S3	9	10	10	10	10	10	10	10	10	10	10	9.91
S4	8	9	10	9	10	10	9	10	9	10	10	9.45
S5	10	10	10	10	10	10	9	10	10	10	10	9.91
S6	10	10	10	10	10	10	10	10	10	10	10	10.00
S7	10	10	10	10	10	10	10	10	10	10	10	10.00
AVE	9.29	9.71	10.00	9.71	10.00	10.00	9.57	10.00	9.71	10.00	10.00	**9.82**

(4) Group B responses to utterance of s3202b, 291.988 sec (control)
(rate = 9.18 syl/sec, reduction = 0.36)

B5	there's	no	reason	to	be	scared	of	anything	that	you	do	AVE
S1	9	10	10	10	10	10	10	10	10	9	10	9.82
S2	10	10	10	10	10	10	9	10	10	10	10	9.91
S3	8	9	10	10	10	10	10	10	10	10	10	9.73
S4	10	10	10	9	10	10	9	10	10	10	10	9.82
S5	9	10	10	10	9	10	10	10	10	10	10	9.82
S6	9	10	10	10	10	10	10	10	10	10	10	9.91
AVE	9.17	9.83	10.00	9.83	9.83	10.00	9.67	10.00	10.00	9.83	10.00	**9.83**

The small difference between the average scores of Group A and Group B is not statistically significant ($p > 0.9$). Therefore, it may be concluded that Group A and Group B performed very similarly on the control utterance, indicating that they are equally matched.

3.2. Natural Fast Utterances vs. Time-Compressed Slow Utterances

Natural fast utterances and time-compressed slow utterances show intelligibility differences in both directions. There are three utterance pairs in which the natural fast utterance appears to be more easily perceived than the corresponding time-compressed slow utterance. These data are

presented below in tables (5)-(10).

(5) Group A responses to natural fast utterance of s1003a, 343.263 sec (rate = 10.51 syl/sec, reduction = 0.40)

A5	and	I	go	why	are	you	buying	that	and	she	goes	AVE
S1	0	9	10	8	0	10	9	10	0	9	10	6.82
S2	7	8	10	10	9	10	10	10	9	10	10	9.36
S3	0	10	8	9	0	9	10	10	0	10	10	6.91
S4	2	3	4	7	6	7	8	10	9	10	10	6.91
S5	8	9	10	10	9	10	10	10	9	10	10	9.55
S6	3	10	10	10	0	10	10	10	9	10	10	8.36
S7	8	9	10	10	9	10	10	10	9	10	10	9.55
AVE	4.00	8.29	8.86	9.14	4.71	9.43	9.57	10.00	6.43	9.86	10.00	**8.21**

(6) Group B responses to corresponding time-compressed slow utterance

B6(x)	and	I	go	why	are	you	buying	that	and	she	goes	AVE
S1	10	10	0	0	0	8	0	9	8	9	10	5.82
S2	10	10	9	10	9	10	10	10	10	10	10	9.82
S3	10	10	0	0	0	8	0	10	9	10	10	6.09
S4	10	10	0	0	0	0	0	0	0	9	10	3.55
S5	9	10	0	10	0	10	0	0	0	0	0	3.55
S6	9	10	0	8	0	10	0	10	8	10	10	6.82
AVE	9.67	10.00	1.50	4.67	1.50	7.67	1.67	6.50	5.83	8.00	8.33	**5.94**

(7) Group A responses to natural fast utterance of s0305, 361.789 sec (rate = 9.36 syl/sec, reduction = 0.46)

A12	he	was	telling	me	a	little	bit	about	it	and	I	said	AVE
S1	10	10	10	10	10	10	10	10	0	10	10	10	9.17
S2	10	10	10	10	10	10	10	10	10	9	10	10	9.92
S3	10	10	10	10	9	10	10	10	0	9	10	10	9.00
S4	10	0	10	10	9	10	10	10	0	0	9	10	7.33
S5	8	0	10	10	9	10	10	10	8	9	10	10	8.67
S6	9	10	10	10	10	10	10	10	8	9	10	10	9.67
S7	7	8	10	10	10	10	10	10	7	9	10	10	9.25
AVE	9.14	6.86	10.00	10.00	9.57	10.00	10.00	10.00	4.71	7.86	9.86	10.00	**9.00**

(8) Group B responses to corresponding time-compressed slow utterance

B8(x)	He	was	telling	me	a	little	bit	about	it	and	I	said	AVE
S1	9	10	10	10	10	9	10	10	0	9	10	10	8.92
S2	10	10	10	10	9	10	10	10	10	9	10	10	9.83
S3	10	10	10	10	9	10	10	10	0	10	10	10	9.08
S4	10	10	10	10	9	10	10	10	10	0	9	10	9.00
S5	10	10	10	10	9	10	10	10	0	0	0	0	6.58
S6	10	10	10	10	10	9	10	10	10	0	9	10	9.00
AVE	9.83	10.00	10.00	10.00	9.33	9.67	10.00	10.00	5.00	4.67	8.00	8.33	**8.74**

(9) Group A responses to natural fast utterance of s1002a, 149.937 sec
(rate = 9.50 syl/sec, reduction = 0.49)

A9	ten	thirty	in	the	morning	so	I	can	go	to	AVE
S1	10	10	10	10	10	8	9	9	10	10	9.60
S2	10	10	10	10	10	7	8	9	10	10	9.40
S3	10	10	10	10	10	6	7	8	9	0	8.00
S4	10	10	10	10	10	0	10	0	10	10	8.00
S5	10	10	10	10	10	7	8	9	10	10	9.40
S6	10	10	10	10	10	7	8	9	10	10	9.40
S7	10	10	10	10	10	10	10	9	10	10	9.90
AVE	10.00	10.00	10.00	10.00	10.00	6.43	8.57	7.57	9.86	8.57	**9.10**

(10) Group B responses to corresponding time-compressed slow utterance

B11(x)	ten	thirty	in	the	morning	so	I	can	go	to	AVE
S1	10	10	10	10	10	10	10	10	10	10	10.00
S2	10	10	10	10	10	10	10	10	10	10	10.00
S3	10	10	10	10	10	7	8	9	10	10	9.40
S4	10	10	10	10	10	0	0	0	0	0	5.00
S5	9	10	10	10	10	0	8	0	0	10	6.70
S6	10	10	8	9	10	9	10	10	10	10	9.60
AVE	9.83	10.00	9.67	9.83	10.00	6.00	7.67	6.50	6.67	8.33	**8.45**

Comparison of the pairs in tables (5)-(6), (7)-(8), and (9)-(10) suggest that the natural fast utterance in each case is more intelligible than the time-compressed slow utterance—more significantly in the case of (5)-(6) than in (7)-(8) or (9)-(10). However, none of these differences is found to be significant ($p > 0.06$ for (5)-(6); $p > 0.6$ for (7)-(8); $p > 0.5$ for (9)-(10)).

In the remaining four utterance pairs, the time-compressed slow utterance appears to be more easily perceived than the natural fast utterance. These data are presented below in tables (11)-(18).

(11) Group A responses to natural fast utterance of s3202b, 353.036 sec
(rate = 9.15 syl/sec, reduction = 0.44)

A2	and	if	I	continue	to	do	that	for	a	long	AVE
S1	0	0	0	0	0	0	0	0	0	0	0.00
S2	6	7	8	0	0	0	0	8	9	10	4.80
S3	0	0	0	0	0	0	0	0	0	0	0.00
S4	10	0	0	0	0	0	0	0	0	0	1.00
S5	7	8	9	0	9	9	9	0	9	10	7.00
S6	0	3	4	5	6	7	8	9	10	10	6.20
S7	10	8	9	0	0	0	0	8	9	10	5.40
AVE	4.71	3.71	4.29	0.71	2.14	2.29	2.43	3.57	5.29	5.71	**3.49**

(12) Group B responses to corresponding time-compressed slow utterance

B3(x)	and	if	I	continue	to	do	that	for	a	long	AVE
S1	10	10	10	10	10	10	10	10	10	10	10.00
S2	10	10	10	10	10	10	10	10	10	10	10.00
S3	10	10	10	10	10	10	10	0	0	0	7.00
S4	10	9	10	10	10	10	10	8	9	10	9.60
S5	9	10	10	10	10	10	10	10	10	10	9.90
S6	10	10	10	10	10	10	10	0	0	10	8.00
AVE	9.83	9.83	10.00	10.00	10.00	10.00	10.00	6.33	6.50	8.33	**9.08**

(13) Group B responses to natural fast utterance of s1004, 84.831 sec
(rate = 9.27 syl/sec, reduction = 0.58)

B2	are	much	better	than	look	in	the	mirror	and	going	AVE
S1	8	10	10	10	10	0	0	0	0	0	4.80
S2	0	10	10	10	10	9	10	10	10	10	8.90
S3	10	10	10	10	10	8	9	10	10	10	9.70
S4	9	10	10	10	10	8	9	10	10	10	9.60
S5	9	10	10	10	10	8	9	10	9	10	9.50
S6	0	10	10	10	10	8	9	10	9	10	8.60
AVE	6.00	10.00	10.00	10.00	10.00	6.83	7.67	8.33	8.00	8.33	**8.52**

(14) Group A responses to corresponding time-compressed slow utterance

A3(x)	are	much	better	than	look	in	the	mirror	and	going	AVE
S1	9	10	10	10	9	10	10	10	10	10	9.80
S2	10	10	10	10	8	9	10	10	10	10	9.70
S3	10	8	0	10	0	8	9	10	10	10	7.50
S4	10	10	10	10	9	9	10	10	10	10	9.80
S5	10	10	10	10	10	9	9	10	9	10	9.70
S6	10	10	10	10	10	10	10	10	10	10	10.00
S7	10	10	10	10	10	9	9	10	10	10	9.80
AVE	9.86	9.71	8.57	10.00	8.00	9.14	9.57	10.00	9.86	10.00	**9.47**

(15) Group B responses to natural fast utterance of s3202b, 287.858 sec
(rate = 9.12 syl/sec, reduction = 0.43)

B9	if	you	know	how	to	put	an	operating	system	on	your	machine	AVE
S1	10	9	10	10	10	7	9	9	10	10	10	10	9.50
S2	10	8	10	10	10	10	8	9	10	9	10	10	9.50
S3	10	9	10	10	10	10	6	7	10	9	10	10	9.25
S4	9	8	9	10	9	4	5	6	7	8	9	10	7.83
S5	10	9	10	10	10	7	8	9	10	10	10	10	9.42
S6	10	8	10	10	10	10	9	0	10	9	10	10	8.83
AVE	9.83	8.50	9.83	10.00	9.83	8.00	7.50	6.67	9.50	9.17	9.83	10.00	**9.06**

(16) Group A responses to corresponding time-compressed slow utterance

A8(x)	if	you	know	how	to	put	an	operating	system	on	your	machine	AVE
S1	10	10	10	10	10	10	0	10	10	10	10	10	9.17
S2	10	10	10	10	10	10	9	10	10	10	10	10	9.92
S3	9	10	10	10	10	10	0	10	10	10	10	10	9.08
S4	10	10	10	10	10	10	9	10	10	10	10	10	9.92
S5	10	10	10	10	10	10	0	10	10	10	10	10	9.17
S6	10	10	10	10	10	9	0	10	10	10	10	10	9.08
S7	10	10	10	10	10	9	9	10	10	10	10	10	9.83
AVE	9.86	10.00	10.00	10.00	10.00	9.71	3.86	10.00	10.00	10.00	10.00	10.00	**9.45**

(17) Group B responses to natural fast utterance of s1201a, 191.817 seconds
(rate = 9.70 syl/sec, reduction = 0.49)

B12	and	all	this	stuff	and	I	said	wait	a	minute	AVE
S1	0	9	10	0	0	0	0	10	0	0	2.90
S2	0	10	10	10	0	0	0	0	0	0	3.00
S3	0	10	10	8	9	0	0	0	0	0	3.70
S4	0	8	10	0	0	0	0	0	0	0	1.80
S5	0	10	10	0	0	0	0	0	0	0	2.00
S6	0	10	10	0	0	0	0	10	0	0	3.00
AVE	0.00	9.50	10.00	3.00	1.50	0.00	0.00	3.33	0.00	0.00	**2.73**

(18) Group A responses to corresponding time-compressed slow utterance

A11(x)	and	all	this	stuff	and	I	said	wait	a	minute	AVE
S1	10	10	10	10	10	10	10	0	0	0	7.00
S2	10	10	10	10	9	10	10	10	10	10	9.90
S3	10	10	10	10	10	10	10	0	0	0	7.00
S4	10	10	10	10	10	10	10	0	0	0	7.00
S5	10	10	10	10	10	10	10	0	0	0	7.00
S6	10	10	10	10	10	10	10	8	9	10	9.70
S7	10	9	10	9	10	10	10	10	9	10	9.70
AVE	10.00	9.86	10.00	9.86	9.86	10.00	10.00	4.00	4.00	4.29	**8.19**

Comparison of the pairs in tables (11)-(12), (13)-(14), (15)-(16), and (17)-(18) show that the time-compressed slow utterance in these cases is more intelligible than the natural fast utterance.

While the differences in (13)-(14) and (15)-(16) are not found to be significant ($p > 0.2$ in both cases), the differences in (11)-(12) and (17)-(18) are found to be significant ($p < 0.003$ for (11)-(12); $p = 0$ for (17)-(18)).

4. Discussion

Some methodological points should be addressed before discussion of the results. First, splicing of the speech samples into gates was very difficult due to the high degree of coarticulation between segments in fast speech; thus, there were some gates in which the beginning of a cue for the first segment of the following word could not be cut out because it was overlaid on the last segment of the final word of the gate. Second, the utterances extracted from the corpus, while similar in the number of words they contained, differed from each other on at least two other dimensions: (i) whether they were complete thoughts as opposed to fragments, and (ii) how many content words vs. function words they contained. In addition, the quality of the speech samples themselves varied in a number of ways: the level of background noise, the amplitude of the signal (e.g., some talkers spoke more quietly), and the voice quality of the talker (e.g., one talker spoke in a very creaky voice), to name a few. It should also be noted that, due to the incremental nature of the experimental design, top-down information likely figured heavily into subjects' responses. As noted above, due to the nature of the task, the subjects that listened to the natural fast utterance and the subjects that listened to the time-compressed slow utterance in a given pair were not the same; they were deliberately kept different in order to circumvent the priming effect that would have manifested itself if the same group had listened to both utterances in the pair (which, consequently, would have made it impossible to compare the intelligibility of the two utterances in the pair). Finally, the small number of subjects did not allow for many statistically significant results.

When measures were available to equalize these disparities, they were taken. For example, with regard to the varying amplitude of the samples, an effort was made to make them more equal by manipulating the volume at which the samples were played; thus, softer samples were played at a constant higher volume than louder samples. The influence of most other inequalities between utterance pairs, as well as the role of top-down information, may be assumed to have been more or less canceled out by the fact that these factors should have affected both subject groups equally, since all subjects heard the same sequences of words in the utterances of interest. It is acknowledged, however, that the gating of the time-compressed slow speech might have been cleaner than that of the natural fast speech due to the lower degree of coarticulation between word-edge segments in slow speech. Both this aspect of the speech samples played to listeners and the more "natural" voice quality of the natural fast utterances may have had an effect on subjects' pattern of responses between the two groups. In a sense, though, it would actually undermine the study to eliminate these differences, since these are probably among the very factors underlying a difference in intelligibility between natural fast utterances and time-compressed slow utterances.

Overall, the results obtained in this study are consistent with those of Janse (2004): artificially time-compressed speech is easier to process than naturally fast speech. While most of the results obtained were not statistically significant, those that were statistically significant showed that time-compressed slow speech is indeed easier to perceive than natural fast speech. It appears, then, that Janse (2003) is correct in claiming that "changes in temporal pattern and in segmental intelligibility that accompany fast speech do not occur because speakers want to help their listeners, but rather because speakers cannot speed up otherwise" (123).

5. Conclusion

The data collected in this study support the findings of Janse (2004) that artificially time-compressed speech is easier to process than naturally fast speech. This corroboration may be taken

to reaffirm the importance of prosody in speech processing, as prosody appears to underlie the opposite result of Janse (2003) with synthetic diphone speech (cf. §1.2). The unnatural prosody of synthetic speech—here, the lack of prosodic alternation, peaks and valleys, which can serve as cues to word boundaries—results in the lower intelligibility of synthetic speech, and this is what Janse herself suggests. The degrading effect of unnatural prosody on intelligibility may also account for the complementary findings of Pickett and Pollack (1963) for temporally stretched speech (cf. §1.2). In this case, the altered prosody of the fast speech rate creates a clash when laid over a normal or slow speech rate; this mismatch is also seen to decrease intelligibility.

A final point to take away from these results is that articulatory cues always seem to be favorable for speech intelligibility. Rather than balking at the task of processing temporally dense clusters of cues in hyperarticulated fast speech, listeners appear to make use of whatever cues they can get. The lower gap in figure (2) above, then, is not due to a deliberate communicative empathy-motivated pattern on the part of speakers, but rather to simple human articulatory limitations.

References

Janse, Esther. (2003). *Production and Perception of Fast Speech*. PhD dissertation, Utrecht University. Utrecht, The Netherlands: Landelijke Onderzoekschool Taalwetenschap (LOT) Dissertation Series 69. Available online: http://www.let.uu.nl/~Esther.Janse/personal/dissertatieEJanse.pdf.

---. (2004). "Word perception in fast speech: artificially time-compressed vs. naturally produced fast speech." *Speech Communication* 42: 155-173.

Lindblom, Björn. (1990). "On the Communication Process: Speaker-Listener Interaction and the Development of Speech." *AAC Augmentative and Alternative Communication*: 220-230.

Pickett, J.M. and Irwin Pollack. (1963). "Intelligibility of excerpts from fluent speech: Effects of rate of utterance and duration of excerpt." *Language and Speech* 6: 151-164.

Shockey, Linda. (1998). "Perception of reduced forms by non-native speakers of English." In D. Duez (ed.), *Sound Patterns of Spontaneous Speech*. Aix: ESCA. 97-100.

Soltau, Hagen and Alex Waibel. (1998). "On the influence of hyperarticulated speech on recognition performance." In *Proceedings of the 5th International Conference on Spoken Language Processing* (ICSLP): paper 0736.