Articulatory Correlates of Prosodic Control: Emotion and Emphasis *

DONNA ERICKSON, OSAMU FUJIMURA, and BRYAN PARDO

The Ohio State University

KEY WORDS

XXX

XXX

XXX

XXX

ease provide

ABSTRACT

This study examines mandibular correlates of prosodic control in nonread dialog exchanges, in which the subject is asked to repeat the same correction of one digit in a three-digit sequence consisting of "five" or "nine" followed by "Pine Street." Articulatory and acoustic data were collected for four speakers of American English at the X-ray Microbeam Facilities at the University of Wisconsin. Jaw opening was measured as vertical jaw position at the time of maximum opening. Middle digits perceived by independent listeners as emphasized generally show jaw opening which is larger than the average jaw opening for the utterances in which they occur. As the speaker repeatedly makes the same correction, not only does jaw opening increase significantly on the corrected digit but also the overall amount of jaw opening on all digits in the corrected exchanges increases. Independent separate perception tests show that listeners also perceive the speakers' answers to be more irritated as the speaker repeats the same correction. The findings suggest a local and global use of the jaw opening gesture to produce both linguistic or paralinguistic and extralinguistic information, that is, word emphasis and the emotional tenor of the dialog itself.

INTRODUCTION

Recent studies of articulatory (in particular, mandible) correlates of prosodic control have focused on read-from-text speech (Beckman & Edwards, 1992; Beckman, Edwards, & Fletcher, 1992; Cohen, Beckman, Edwards, & Fourakis, 1995; Edwards, Beckman, & Fletcher, 1991; Erickson, Lenzo, & Fujimura, 1994; Erickson & Fujimura, 1996a; Fox & Josephson, 1992; Fox, Josephson, & Erickson, 1991; Harrington, Fletcher, & Beckman, in

* Acknowledgments: Work supported by NSF SBR-951199B, ATR/HIP, and from a collaborative support given to Osamu Fujimura from ATR/MIC, Kyoto, Japan (President: Ryohei Nakatsu). We also thank Yoh'ichi Tohkura and Kiyoshi Honda at ATR/HIP, Kyoto, Japan for their support; part of the work in earlier phases was conducted by two of the coauthors as guest researchers at ATR. We also acknowledge our gratitude for the support of Rob Fox, Chair of the Department of Speech and Hearing Science, and Peter Culicover, Director of the Center for Cognitive Science, The Ohio State University. In addition, we thank Jun-Ping Gong for technical help, Rob Leighty for statistical consulting, John Westbury and the Microbeam Facility staff at the University of Wisconsin for making their facility and support available for data acquisition, and Deirdre Smith for her support. We also thank the two reviewers, Gerry Docherty and Ailbhe Ni Chasaide, for their helpful suggestions that resulted in a substantive improvement of the paper.

Address for correspondence: Donna Erickson, 110 Pressey Hall, 1070 Carmack Road, The Ohio State University, Columbus, Ohio 43210. E-mail: <erickson@shs.ohio-state.edu>

press; Harrington, Palethorpe, Fletcher, & Beckman, 1996; de Jong, 1995; de Jong, Beckman, & Edwards, 1993; Macchi, 1985; Oshima & Graceo, 1993; Summers, 1987; Westbury & Fujimura, 1989). With regard to analysis of jaw movement in connection with contrastive emphasis, increased jaw opening has been shown to occur on emphasized syllables in read speech (e.g., Erickson, to appear; Erickson & Fujimura, 1996a; de Jong, 1995; de Jong et al., 1993; Fujimura, 1990; Westbury & Fujimura, 1989). The finding of increased jaw opening for emphasized syllables is compatible with reported results of acoustic studies that associate increased duration and intensity with emphasis (e.g., Cooper, Eady, & Mueller, 1985; Lehiste, 1970), since larger jaw opening presumably would be mechanically related to both louder and longer utterances (e.g., Schulman, 1989).

The topic of the relation between acoustics and articulation is not addressed in this paper. The motivation for examining jaw movement independent from acoustic characteristics of emphasis derives from studies which suggest that increased jaw displacement might be correlated with an overall greater degree of effort involved in the production of emphasis (see e.g., Harrington et al., in press; de Jong, 1995; Lindblom, 1990). Also, the underlying rhythmical structure of utterances may be related to jaw opening patterns (e.g., Erickson, to appear; Erickson & Fujimura, 1996b; Fowler, 1983; Fujimura & Erickson, 1996; MacNeilage, in press; Patel, Löfqvist, & Naito, in press; Tuller & Fowler, 1980; Vatikiotis-Bateson & Kelso, 1992).

Along these lines, Fujimura (1992, 1994, in press) suggests that the prosodic characteristics of an utterance can be described in terms of a magnitude distribution pattern of a series of pulses representing syllables and boundaries. According to the C/D model proposed by Fujimura, there is a certain prescribed relation between syllable magnitudes and syllable durations for a given utterance condition; the syllable magnitude distribution is determined by the prosodic specification of the utterance including contrastive emphasis of a part of the sentence. It is suggested by Erickson (to appear), Erickson and Fujimura (1996a) and Fujimura and Erickson (1996), based on analysis of read speech, that an approximation of the syllable magnitudes (and syllable durations) may be related to the amount of jaw opening associated with each syllable.

It is known that "read-from-text" prosody differs considerably from "nonread" dialog exchanges in terms of rhythm, tempo, pauses, F0 patterns, etc. (e.g., Barick, 1979; Beckman, 1995; Goldman-Eisler, 1968). Presumably in dialog, speakers use prosody to interactively convey both linguistic and paralinguistic information to listeners. For instance, speakers use pauses to convey how a sentence is grammatically-parsed, as well as to signal turn-taking information. F0, among other acoustic properties, is used to signal word emphasis. On the other hand, a speaker might modify the F0 pattern of the sentence to reduce or reassign prominence, if s/he thought the listener already knew parts of the message. In addition, the emotional demands of the dialog situation may affect prosody: F0 and intensity are known to increase with anger, for example, (Leinonen, Hiltunen, Linnankoski, & Laakso, 1997; Scherer, 1986; Williams & Stevens, 1972). The interaction of emotion with articulation, specifically jaw movement, and with emphasis has not been studied to date.

The general interest of this study is to examine the effects on articulation and their interactions with linguistic (such as vowel articulation), paralinguistic (such as focus), and extralinguistic (such as speaker's irritation) factors of the message. The data pertain to

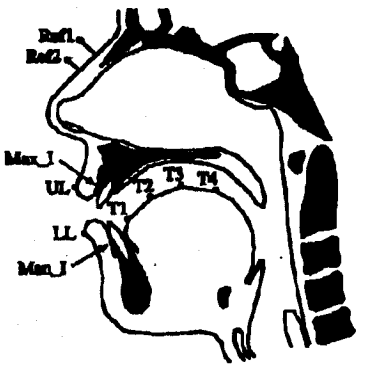

Figure 1

Placement of pellets on tongue, lips, jaw, and reference points.

speakers' corrections of the partner's responses in "nonread" dialog exchanges. The following questions are addressed: (1) Does jaw opening increase for emphasis placed on a focused word in dialog situations?; and (2) How is jaw opening affected as the speaker is asked to repeat the same correction many times within a single dialog exchange? Presumably this task would not only introduce focus and emphasis within the dialog, but may also evoke the speaker's irritation, anger, frustration, or other emotions.

METHODS

Data recordings

In order to address these questions, we examined the articulatory correlates of repeated corrections in simulated conversations which were conducted as a (nonread) dialog between the subject and the experimenter in a recording set up at the X-ray Microbeam Facilities at the University of Wisconsin. (For a description of the microbeam method, see Fujimura, Ishida, & Kiritani, 1973; Kiritani, Itoh, & Fujimura, 1975; Nadler, Abbs, & Fujimura, 1987; Westbury, 1994; Westbury, Milenkovic, Weismer, & Kent, 1990.) This facility allows the user to track the motion of a speaker's articulators as s/he talks by tracking gold pellets affixed to various articulators. Spherical gold pellets (2.5–3 mm in diameter) were affixed to the tongue, lips, and jaw of the speakers (see Figure 1). Two pellets were attached to the mandible, one at the incisors, and another on a molar tooth (not shown in Figure 1), and they were sampled at a rate of 80 samples/sec (or 40 samples /sec for Speaker 2). For this study, the pellet attached to the point midway between the central mandibular incisors was used for the analysis of the vertical movement of the jaw. In addition, reference pellets were affixed midsagitally to the nose bridge and to the anterior surface of the midpoint between the central maxillary incisors. These references were used to correct for head movement during the utterance to produce pellet position data fixed relative to the maxillary occlusal plane (see Westbury, 1991, for a detailed description). The x-axis of the data corresponds to the intersection of the midsagittal plane and the maxillary occlusal plane, and the origin of the coordinate system corresponds to the lowermost edge of the maxillary

incisor. The y-axis is normal to the maxillary occlusal plane, intersecting the plane at the origin. The y-coordinate value of the mandibular incisor pellet (henceforth jaw pellet) represents the vertical distance from the maxillary occlusal plane to the center of the pellet sphere attached to the mandible incisor, and it was always negative.

Speech samples

Between 38 and 70 target phrases were elicited from each of four speakers of Midwestern American English (two men, two women). The target phrase was always one of the following three digit sequences: "5 9 5," "9 5 9," or "5 5 9." The elicitation scenario called for a correction of one of the three digits by replacing "five" by "nine" or vice versa in the experimenter's question utterance (see Example 1). This correction occurred in the first, second, or the last of the three digits in the sequence. The subject responding to the experimenter's question repeated the same correction up to five or six times. The entire dialog, terminated automatically at 25 seconds, was recorded continuously with the microbeam pellet tracking. The experiment elicited 12 to 18 dialogs from each of the subjects.

The speakers were instructed by the experimenter to pretend the interaction was a telephone conversation and to reply to the questions initially for each dialog by reading the prompt on the monitor. If the elicitor indicated she was having problems hearing the response clearly, they were also told "not to read the prompt in the monitor screen but to try to get the correct information across according to what the monitor specified in the beginning of the dialog." The elicitor sat out of sight but within hearing distance of the speaker. There was no monitor-displayed text prompting the speaker after the initial answer (hereafter "reference utterance") had been spoken. In order to evoke emotional (irritated) responses from the speaker, the experimenter maintained a formal clinical style of interaction throughout the experiment, including instruction-giving and signing of the human subjects form. The experimenter wore a white lab coat, combed her hair back to enhance a serious attitude, minimized personal small-talk, and spoke in a clinical "no-nonsense" tone of voice with moderate change in pitch range.

The series of exchanges between elicitor and speaker is referred to as a "dialog set." It always includes one reference utterance mentioning a three digit address, followed by several attempts to correct the elicitor's "misunderstanding" of one digit in the address mentioned in the reference utterance. In each dialog set, the elicitor always "misunderstands" the same digit in the address. A typical dialog is given below. This dialog set includes four corrections. The first exchange is the answer by the speaker to the first question and is referred to as the "reference utterance" (indicated in italics in Example 1). The subsequent exchanges are referred to as the exchange numbers 2 through 5. In this dialog set, the speaker is responding to the elicitor's "misunderstanding" of the middle digit.

Since the speakers were not reading the responses, but engaging in dialog with the experimenter, the verbal and paralinguistic parameters of their responses varied. The speakers responses usually consisted of the three-digits, often including "Pine Street." Sometimes their responses were preceded by such phrases as "No, I'm saying...," "I said...," or "No, you're wrong, it's ..." Occasionally, they would insert sentences such as "No, forget the 9, it's 5 5...and the last number is a 9," "No, there's two 9's and one 5—9 5 9 Pine Street," "Not all 5's, the last one is a 9." In addition, there were disfluencies,

(1) **Dialog 13** (Speaker 2)

1. ELICITOR: Where do you work?
   SPEAKER 2: *I work at 9 5 9 Pine Street.*
2. ELICITOR: I'm sorry, was that 9 9 9 Pine Street?
   SPEAKER 2: No, it's 9 FIVE 9 Pine Street.
3. ELICITOR: Listen, is it 9 9 9 Pine Street?
   SPEAKER 2: It's 9 FIVE 9 Pine Street.
4. ELICITOR: I'm sorry. It's not coming through. Is it 9 9 9 Pine Street?
   SPEAKER 2: No, it's 9 FIVE 9 Pine Street.
5. ELICITOR: You're saying 9 9 9 Pine Street, right?
   SPEAKER 2: No, I'm saying 9 FIVE 9 Pine Street.

such as "No, I haven't been, it's been ni.....5 5 9 Pine Street," or laughing in the responses. Speaker 1 frequently ended her responses with a rising F0 pattern, similar to the question intonation. In addition, the speakers would make frequent pauses, and vary the phrasing, as well as loudness and rate of their responses, often speeding up or slowing down within the same response.

Analysis method

Measurements of the lowest vertical position of the jaw were made for each of the digits using a MATLAB-based software program (Ubedit) developed by the third author. An appropriate instant for onset and offset of the jaw opening pellet tracing corresponding to each of the digits was marked manually, and the lowest vertical jaw position for each digit was calculated automatically by Ubedit. A sample of Ubedit's display of the acoustic wave form and vertical movement of the jaw for Example 1 is shown in Figure 2. Jaw displacement is measured in micrometers as the distance from the maxillary occlusal plane to the maximum opening during the vocalic gesture for each syllable. The time scale is in millimeters.

In the data analysis reported here, only the dialogs eliciting correction on the second (i.e., middle) digit of the three-digit sequence were used (as shown in the sample dialog above). The number of exchanges per dialog differed among speakers due to instances of mistracking of the jaw pellet as well as differences in how the dialogs proceeded. See Table 1 for details on the makeup of the corpus of data used.

Throughout the analysis, the digits "five" and "nine" were treated as being interchangeable, since the data set was small, both contain the same vowel (diphthong), and statistical analysis of the corpus used in the study showed no significant difference between the amount of jaw opening for "five" and "nine." This finding was supported by previous work for read speech which also showed no statistically significant difference between the amount of jaw opening for "five" and "nine" (Erickson & Fujimura, 1996a).

The range of amount of jaw opening varies across speakers, and the jaw opening maxima, minima, mean, and range for each speaker is shown in Table 2 below.

In order to compare jaw opening among different exchanges, dialogs, and speakers, two different types of normalization were used (see examples in Table 3). Under "Exchange

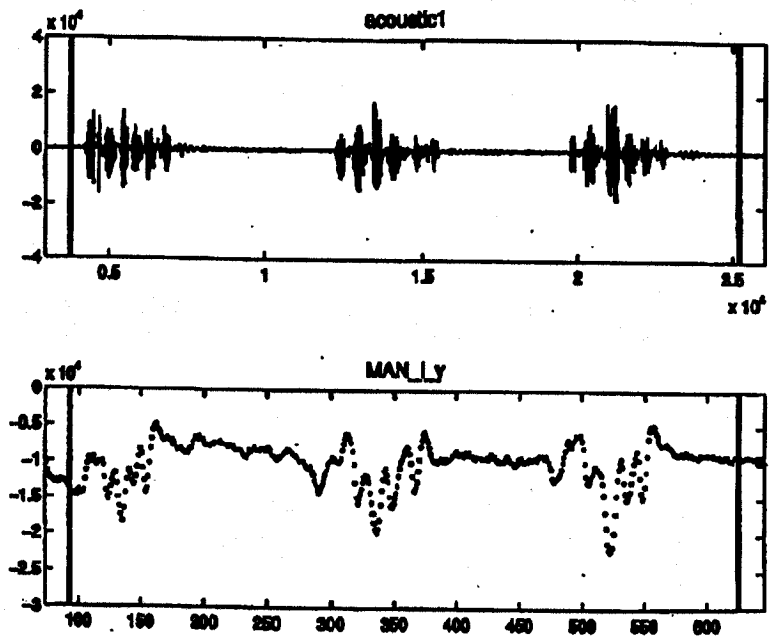

Figure 2

Display of jaw tracing (lower panel) and acoustic signal (upper panel) for first three responses of Speaker 2 to the elicitor's question, "Where do you work?" in Dialog 13.

Normalization," the jaw opening values for the three digits in a single utterance (exchange) were divided by the mean jaw opening for that three digit sequence. The resulting values are in relation to the mean, and thus have no units. The average jaw opening for the three digits in each utterance was also recorded. Since different speakers have different means of jaw opening, average jaw opening in each exchange proved difficult to compare across speakers. This problem was addressed by normalizing exchange average jaw openings for each speaker. Under *speaker-normalized exchange average*, the jaw opening values were normalized by dividing the *exchange average* jaw opening for a particular speaker for each utterance by the *mean exchange average* jaw opening over all (recorded) dialogs of that speaker. The result is a proportion in relation to the average for that speaker.

In the analysis of the results, the convention is followed that in reporting on jaw opening averaged over speakers, the normalized values are used, that is, for the figures and statistical analyses. The raw jaw opening values for individual speakers are reported in tables.

TABLE 1

Data base examined

Speaker	*Exchanges per dialog*	*Number of dialogs*	*Total number of utterances*	*Total number of reference utterances*	*Total number of 1st corrections*	*Total number of 2nd corrections*	*Total number of 3rd corrections*	*Total number of 4th corrections*
1	3	4	12	4	4	4	0	0
2	4	5	20	5	5	5	5	0
3	5	6	30	6	6	6	6	6
4	5	7	35	7	7	7	7	7
Total	17	22	97	22	22	22	18	13

TABLE 2

Range of jaw opening (mm)

Speaker	*Maxima*	*Minima*	*Mean*	*Range of jaw opening*
1	18.17	11.33	14.60	6.84
2	28.48	12.45	17.52	16.03
3	18.01	9.06	13.92	8.95
4	19.76	8.56	13.14	11.20

TABLE 3

Example normalization values for a dialog with three exchanges

	Raw Data (in mm)	*Exchange normalized*	*Exchange Average (in mm)*	*Speaker normalized exchange average*
Reference	20,16,14	1.20, .96, .84	16.66	0.93
2nd Exchange	21,18,15	1.70,1.00,.83	18.00	1.00
3rd Exchange	21,21,16	1.09,1.09,.83	19.33	1.07

Perception tests

The design of the experiment called for speakers to correct street addresses which were misunderstood by the experimenter. The point of this was to elicit emphasis on the corrected digits. Although the speakers did try to correct the misunderstanding of the street address, they did not always put emphasis on the digit intended by the experimental design to be corrected. Emphasis-perception tests were conducted to assess whether listeners actually perceived emphasis on the digit intended by the experimental design to be emphasized. Reported here are perception tests done earlier with the same data, as part of acoustic and articulatory studies by Spring, Erickson, and Call (1992) and Erickson and Lehiste (1995). Perception tests were run on the three digit sequences plus "Pine Street" (excluding the rest of the utterance) uttered by each of the four speakers in separate listening test sessions but with different dialogs mixed. Two randomizations of the sequences were presented to

10 university students. The material included all utterances with intended emphasis placed on initial, middle, and final digits, and reference utterances. For each spoken three digit sequence presented to the listeners, they were asked to circle the digit on the answer sheet they heard as emphasized in the utterance. Only results for utterances with middle digits intended to be emphasized were analyzed and reported in the following section.

Emotion-perception tests were also run for each speaker to see whether listeners perceived irritation as the speakers repeated the same correction. A prior study asked listeners to label the utterances of Speaker 1 with any one-word label describing the emotional state of the speaker. The most common responses were then collated and found to separate naturally into five groups: neutral, emphatic, questioning, happy, and irritated (Spring, Erickson, & Call, 1992). The emotion-perception test for this study used these five groups as the labels by which the listeners had to identify each utterance. For the emotion-perception tests, two randomizations of each speaker's whole responses (digit sequences and surrounding words), were administered to 10 university students. Listeners were asked to circle the label which best described each sentence. Listeners were not allowed to circle more than one response per utterance. Since each of the ten listeners was presented with two randomizations of the same data, there was a total of 20 categorizations of each utterance. The results, as discussed below, focus on the irritation scores, since this was the emotion-category most anticipated from the experimental design.

RESULTS

Emphasis-perception test

A spoken sample of a digit was judged to have INTENDED emphasis if the experimental design called for that digit to be corrected in the speaker's response to the elicitor's misstatement of the address. In this paper, these are always middle digits. A spoken sample was judged to have PERCEIVED emphasis if 80% of the listeners' judgments indicated that digit (initial, middle, or final) as the one which was emphasized in a particular utterance. The cut-off point of 80% was arrived at by searching for that point which would result in the smallest number of samples whose intended and perceived emphasis differed, considering all data of the four speakers together. Table 4 describes the results of the emphasis-perception test. A HIT is a sample with both perceived and intended emphasis (all middle digits). A MISS is a sample with intended but no perceived emphasis (all middle digits). A CORRECT REJECTION (middle digits in a reference utterance or initial and final digits) is a sample with neither intended nor perceived emphasis and a FALSE POSITIVE is a sample with perceived but no intended emphasis (middle digits in a reference utterance or initial and final digits).

Listeners perceived emphasis on the middle digits intended by the experimental design on the average approximately 65% of the time. This result was different from work with contrastive emphasis in read speech, in which listeners were able to perceive emphasis on the intended emphasized digit 100% of the time (Erickson & Fujimura, 1996a). It may be noted that in the read speech Pine Street experiment, each speaker was given a prompt in the text of the sentence to be uttered with the corrected digit printed in capital letters, in contrast to Arabic numerals, and s/he was told that this showed the corrected part in the answer of the utterance. In this sense, in the "read speech," the speaker consciously places emphasis on the corrected digit.

TABLE 4

Results of the emphasis-perception test. (Analysis of only those utterances with intended emphasis on the middle digit)

Speaker	*Total examined*	*Intended emphasis*	*Correct rejections*	*False positive*	*Hits*	*Misses*	*No perception data*	*Hit rate*
1	36	8	28	0	4	4		50%
2	60	15	44	1	11	4		73%
3	90	24	62	2	16	7	3	67%
4	105	28	72	3	18	9	3	64%
Total	291	75	206	6	49	24	6	65%

Jaw opening on emphasized digits

Previous results pertaining to read speech showed the emphasized digit had the greatest mean maximum jaw opening in the utterance (e.g., Erickson, to appear; Erickson & Fujimura, 1996a). This study shows a similar effect in nonread dialog. Figure 3 shows the mean of exchange-normalized values of jaw opening for all four speakers for the first, second, and third digits for those exchanges in which the middle digits were hits (intended and perceived as emphasized). The bars indicate the mean values for the first, second, and third digits in three digit utterances. A value of 1 indicates that the maximum jaw opening for that digit is equal to the average jaw opening for its exchange.

The jaw opening values of the second digit hits is generally greater than the mean of the jaw opening for the other digits in the same exchange. Middle digits average 16.63 mm of jaw opening versus 15.60 mm for first and 13.45 mm for third digits. A *t*-test comparing the middle digits to the initial and final digits combined finds the difference in the means to be significant with a $p < .01$. Table 5 shows the mean jaw opening values together with the exchange-normalized values, for the digits in those exchanges perceived as middle digit emphasized, broken down by speaker. For Speaker 3, although the mean of raw jaw opening values for the middle digit appears smaller than that for the initial digit, in terms of exchange-normalized values, the value for the second digit is greater than unity (1.06>1.00). This shows that on the average, the second digit has a larger value than the mean of all three digits even for this speaker.

Digits which should have been emphasized according to the requirements of the dialog but were not perceived as emphasized (i.e., misses) did not show greater than average jaw opening in their exchanges. Figure 4 displays the mean of the exchange-normalized jaw opening values, averaged over all four speakers, for hits, misses, correct rejections, and false positives. A value of 1 indicates that the maximum jaw opening for that digit is equal to the average jaw opening for its exchange. Bars indicate the mean values for "Hits" (tokens with intended emphasis which were perceived as emphasized by at least 80% of listeners), "Misses" (tokens with intended emphasis which were not perceived as emphasized by at

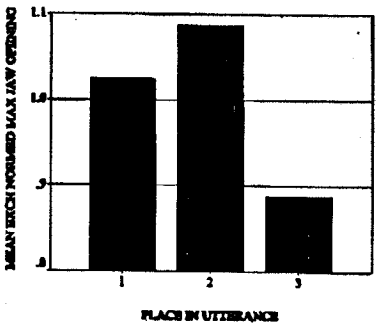

Figure 3

Mean exchange-normalized maximum jaw opening values for four speakers for only those tokens in those utterances in which the middle tokens were "Hits" (tokens with intended emphasis which were perceived as emphasized by at least 80% of listeners).

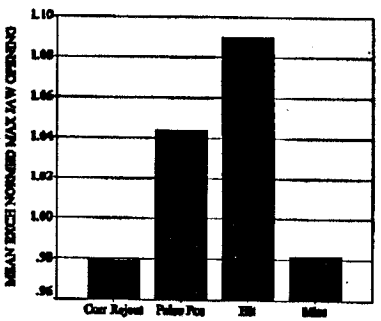

Figure 4

Mean exhange-normalized maximum jaw opening values for all measured tokens in all utterances by all speakers.

least 80% of listeners), "Correct Rejections" (tokens without intended emphasis which were not perceived as emphasized by at least 80% of listeners) and "False Positives" (tokens without intended emphasis which were perceived as emphasized by at least 80% of listeners). As shown in Table 6, the jaw opening for hits averages 16.7 mm, which is 2.6 mm greater than the 14.1 mm average for the misses. Using exchange normalized data for all the speakers, a *t*-test comparing hits and misses indicates that this difference in jaw opening is statistically significant ($p<.01$).

These results suggest that increased jaw opening is an articulatory characteristic of effectively emphasized digits in spontaneous dialog. Hereafter, the term "emphasized digits" will be used to refer to effectively emphasized digits (i.e., "hits").

One further comment about the digits in Table 6 labeled "False positives." These refer to those digits listeners heard as emphasized but not intended by the experimental design to be emphasized. Whether the speakers actually intended to emphasize these particular digits cannot be known at this point. It is interesting to note that the amount of jaw opening on those digits labeled "False positives" for speakers 3 and 4 tend to be larger than these speakers' "Correct rejections." This observation further supports the statement that increased jaw opening is an articulatory characteristic of what is heard as emphasized.

TABLE 5

Mean jaw opening data (in mm) and exchange-normalized values for exchanges perceived as middle digit emphasized. (Numbers in parentheses are standard deviations)

		Raw jaw opening values			*Exchange-normalized values*		
Speaker		*Initial*	*Middle*	*Final*	*Initial*	*Middle*	*Final*
1	(*N*=4)	15.42 (1.71)	16.29 (1.29)	14.84 (1.74)	.99 (.09)	1.05 (.12)	.95 (.08)
2	(*N*=11)	18.27 (3.31)	21.71 (2.86)	15.89 (1.28)	.98 (.11)	1.17 (.12)	.86 (.09)
3	(*N*=16)	15.71 (1.43)	15.33 (1.46)	12.39 (.90)	1.08 (.06)	1.06 (.07)	.86 (.07)
4	(*N*=18)	13.99 (2.27)	14.92 (1.97)	12.7 (1.92)	1.01 (.12)	1.07 (.07)	.92 (.15)
Average	(*N*=51)	15.6 (2.73)	16.63 (3.34)	13.45 (2.04)	1.02 (.11)	1.09 (.10)	.89 (.11)

TABLE 6

Mean jaw opening values (in mm) for each speaker for the correct rejections, false positives, hits, and misses. Numbers in parentheses are standard deviations

Speaker	*Correct rejection*	*False positive*	*Hit*	*Miss*
RAW JAW OPENING VALUES				
1	14.41 (1.93)	0	16.29 (1.29)	14.26 (.82)
2	16.44 (2.61)	13.73	21.71 (2.86)	18.83 (.25)
3	13.49 (2.10)	14.73 (2.17)	15.42 (1.46)	13.48 (1.42)
4	12.71 (2.25)	13.56 (2.38)	14.89 (2.02)	12.5 (2.43)
Average	13.97 (2.64)	13.98 (.88)	16.71 (3.38)	14.14 (2.77)
EXCHANGE-NORMALIZED VALUES				
1	1.00 (.11)	0	1.05 (.12)	.95 (.02)
2	.96 (.11)	.81 (.81)	1.17 (.12)	1.07 (.11)
3	.98 (.13)	1.08 (.05)	1.07 (.07)	.95 (.04)
4	.98 (.13)	1.10 (.22)	1.07 (.07)	.98 (.05)
Average	.98 (.12)	1.00 (.09)	1.09 (.11)	.99 (.06)

Emotion-perception test

The results of the emotion-perception test are displayed as bar graphs in Figure 5. All exchanges (utterances) for all speakers are included. Each exchange received 20 listener categorizations as one of the following groups: "Neutral," "Emphatic," "Happy," "Irritated," or "Questioning."

For the first exchange (the reference exchange), utterances were categorized as neutral an average of 78% of the time. For the second exchange, (i.e., the first correction), they were categorized as emphatic 75% of the time. This was also expected, since the speaker was being put in a situation to correct the middle digit in the utterance. However, for the subsequent exchanges, the number of utterances perceived as irritated increased, while the number of utterances perceived as emphatic decreased. The results are summarized in Table 7 below.

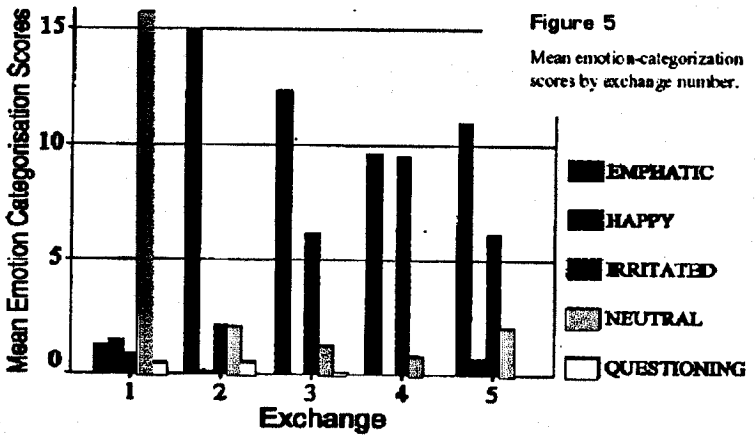

Figure 5

Mean emotion-categorization scores by exchange number.

It is clear from both the figure and the table that there is a trend of increased irritation as exchange number increases. The fifth exchange shows a reduction in mean irritation score in relation to the fourth, but still remains higher than the first two exchanges. Irritation was modeled using a mixed effect ANOVA with *speaker* as a random effect and *exchange* as the fixed effect. The linear orthogonal polynomial component of exchange was significant ($p<.01$) supporting that irritation increases with more exchanges. A bivariate correlation was also performed on exchange and irritation and generated a Pearson Correlation of .51 ($p<.01$). These results indicate that the experiment was generally successful in achieving increasing irritation as speakers were forced to make repeated corrections. Listeners' subjective judgments were based on the whole response of the speaker in each exchange.

Jaw opening on repeatedly corrected middle digits (hits)

Figure 6 shows mean exchange-normalized jaw opening on emphasized middle digits (hits) for the four speakers. All speakers are considered. A value of 1 indicates that the maximum jaw opening for that digit is equal to the average jaw opening for its exchange. The graph indicates that on average the emphasized digit maintains a jaw opening above the average for the exchange in which it happened. Table 8 shows the jaw opening values broken down by speaker, suggesting a general trend for increased jaw opening on the middle digit perceived as emphasized as the speaker repeats the same correction. A *t*-test suggests that the amount of jaw opening of the last correction in the data set analyzed (exchange 2 for Speaker 1, exchange 4 for Speaker 2, and exchange 5 for Speaker 3 and 4) is significantly greater than that of the first correction ($p<.01$). It is interesting to note that for Speaker 1, there were no middle digits perceived as emphasized by listeners for the last intended correction (exchange 3); they were all perceived as not emphasized. A bivariate correlation on all exchanges for all speakers between speaker-normalized exchange average jaw opening and exchange number generates a Pearson Correlation of .59 ($p<.01$). From this it seems there is a significant positive correlation between exchange (i.e., repetition number) and maximum jaw opening on the middle digits perceived as emphasized.

TABLE 7

Results of Emotion-Perception test by speaker and exchange. Values are in terms of mean number of listener-perceptions, with 20 as the maximum score within each exchange

Speaker	*Neutral*	*Emphatic*	*Irritated*	*Happy*	*Questioning*
REFERENCE UTTERANCE					
1	13.25	2.75	.25	1.25	2.5
2	14.8	.4	0	4.4	.4
3	17.5	.67	.83	1	0
4	16.43	1.57	2	0	0
Average	15.5 (78%)	1.35 (7%)	.77 (4%)	1.66 (8%)	.72 (4%)
SECOND EXCHANGE					
1	.5	15.5	1.25	.25	2.5
2	3.6	13.8	1.4	.6	.6
3	1.5	15.33	3.17	0	0
4	2.57	15	2.43	0	0
Average	2.04 (10%)	14.91 (75%)	2.06 (10%)	.21 (1%)	.78 (4%)
THIRD EXCHANGE					
1	1.75	13	5.25	0	0
2	2	12	6	0	0
3	1.5	12.83	5.67	0	0
4	.43	11.86	7.29	0	.43
Average	1.42 (7%)	12.42 (62%)	6.05 (30%)	0 (0%)	.11 (1%)
FOURTH EXCHANGE					
1	–	–	–	–	–
2	0	9.8	10.2	0	0
3	.17	10.67	9.17	0	0
4	2	8.57	9.43	0	0
Average	.72 (4%)	9.68 (48%)	9.6 (48%)	0 (0%)	0 (0%)
FIFTH EXCHANGE					
1	–	–	–	–	–
2	–	–	–	–	–
3	3.17	10.67	6.17	0	0
4	1.14	11.29	6.14	1.43	0
Average	2.16 (11%)	10.98 (55%)	6.16 (31%)	.72 (4%)	0 (0%)

Exchange-average jaw opening

It was thought that repeated corrections of the same information might cause increased irritation on the part of the speaker, and this would be displayed as an increase in the average jaw opening as the exchange number increased. Examination of Figure 7, which displays speaker-normalized jaw opening averaged over all exchanges and dialogs, regardless of the perceived emotion or whether the middle digit was perceived as a hit or a miss, shows

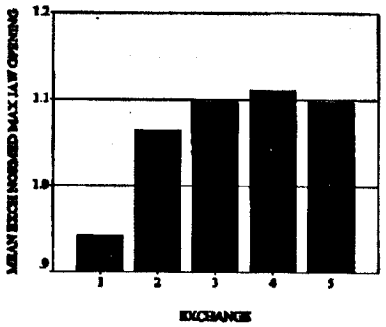

Figure 6

Mean exhange-normalized maximum jaw opening values for middle digits in the initial reference exchanges and middle digit "Hits" (tokens with intended emphasis which were perceived as emphasized by at least 80% of listeners) in exchanges 2 through 5.

TABLE 8

Mean Jaw opening (in mm) for middle digits perceived as emphasized

Speaker	*Exchange 2*	*Exchange 3*	*Exchange 4*	*Exchange 5*
1	16.29 (1.29)	no hits	–	–
2	19.01 (.51)	21.63 (1.73)	24.53 (3.43)	–
3	14.01 (.76)	16.01 (1.17)	15.07 (1.82)	16.35 (.82)
4	14.88 (2.19)	14.43 (1.71)	14.98 (2.97)	15.51 (.67)
Average	15.91 (2.26)	17.46 (3.61)	17.22 (4.83)	15.99 (.83)

that jaw opening increases for each subsequent correction within the corpus, relative to the first correction. The average maximum jaw opening for each individual exchange is calculated by averaging the values for the measured three digit sequence in the exchange. These exchange-averaged jaw openings for each speaker are then normalized to the mean exchange-normalized jaw opening for that speaker. A value of 1 indicates that the average jaw opening in a particular exchange is equal to the average jaw opening across all exchanges for the speaker. The means of these speaker-normalized, exchange-average jaw openings are displayed in the figure, indexed by exchange number. *T*-tests, performed on the exchange average jaw opening normalized over speakers for the second exchange (first correction) versus that in each subsequent correction, suggests that the mean exchange-average jaw openings for each of the exchanges after the first correction are significantly different from that of the first correction, that is, second exchange ($p < .01$).

It is interesting to note that the exchange-average jaw opening follows the same overall pattern of increase which was seen in the irritation scores. The jaw opening values broken down by speaker are shown in Table 9. A bivariate correlation between irritation and speaker-normalized exchange average jaw opening shows significant, though not strong, correlation between the two variables (Pearson Correlation = .35, $p < .01$).

In summary, the following can be said about jaw opening on emphasized digits in dialog situations in which the speaker is asked to repeat the same correction many times: (1) perception of emphasis on a digit tends to occur when it is pronounced with larger jaw

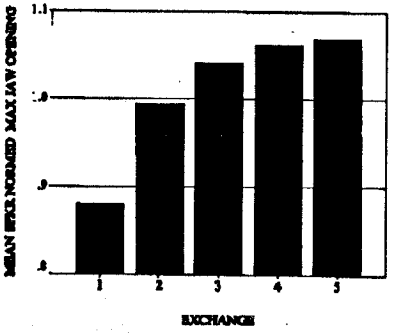

Figure 7

Means of the average maximum jaw opening values for each exchange (utterance) for all exchanges and all speakers.

TABLE 9

Exchange Average Jaw opening (in mm) for each exchange by speaker

Speaker	*Exchange 1*	*Exchange 2*	*Exchange 3*	*Exchange 4*	*Exchange 5*
1	13.34 (.81)	15.52 (.51)	14.95 (.52)	–	–
2	14.91 (.51)	16.94 (.73)	18.75 (1.58)	19.48 (1.07)	–
3	12.01 (.7)	13.69 (.60)	14.64 (1.05)	14.38 (.87)	14.77 (.34)
4	11.65 (1.04)	12.88 (1.66)	13.35 (1.39)	13.75 (1.61)	14.05 (1.63)
Average	12.82 (1.51)	14.50 (1.91)	15.22 (2.36)	15.55 (2.76)	14.38 (1.26)

opening than for the surrounding digits; (2) the maximum jaw opening for both the emphasized digit and its mean for the three digits in the sequence on the average tends to increase as the same digit sequence is repeated several times in response to the elicitor's questions; and (3) as the digit sequence is said repeatedly, independent listeners increasingly categorize the utterances as irritated.

DISCUSSION

The demands of nonread dialog-exchange affect speaker's jaw movement patterns for producing word emphasis. In dialog, increased jaw opening is used to produce emphasis locally on a specific word and also globally on the entire utterance, as the speaker repeats the same correction. The results suggest that the task of making repeated corrections affects the emotional condition of the dialog, for example, an increase in perceived irritation by listeners.

In this paper, the application of increased jaw opening, both locally and globally in response to the repeated requests by the elicitor to correct the same digit, is interpreted as the speaker's response to both the paralinguistic (focus assignment) and extralinguistic (emotional) demands of the dialog task.

The assumption in designing the experiment was that the speaker would produce the target utterances first with no contrastive emphasis, and then, in response to the elicitor's

misunderstanding of one of the digits in the utterance, would produce it with emphasis on the digit that was misunderstood. The results of the emphasis-perception study reported that those digits intended to correct an error were perceived as emphasized only about 65% of the time. In perception studies of read speech, intended emphasis was perceived as emphasized 100% of the time (Erickson & Fujimura, 1996a).

A question arises as to why this study shows a discrepancy between what is intended to be emphasized (by the experimental design) and what is perceived to be emphasized by listeners hearing only the responses of the subjects. Spring, Erickson, and Call (1992) suggest that the extralinguistic emotional demands of the dialog interfere in some way with the perception of emphasis. In their study of a subset of the data for Speaker 1, they report that listeners' perceptions of an emotion like irritation became progressively stronger as the correction was repeated. Furthermore, they report that the utterances listeners identified as most irritated were also the ones in which the listeners had the hardest time identifying the emphasized digit. Spring et al. speculate that the addition of emotion introduced other phonetic information in the form of F0 or loudness, and so forth, which interfered with the perception of emphasis.

However, the hypothesis of interference between emotion and emphasis causing a reduced ability to transmit emphasis is not necessarily supported by the results of this study. Utterances with large irritation scores did not show poor perception of emphasis on the middle digit. In fact, averaged over the four speakers, exchanges containing hits scored an average of 6.12 out of 20 for "irritation" on perception tests while exchanges containing misses averaged only 4.7 out of 20. The mean irritation scores for hits and misses are shown by speaker in Table 10 opposite. Notice for Speaker 1, however, the opposite trend was seen: more exchanges containing misses (26%) were heard as irritated compared to those with hits (6%). For this speaker, emotion may indeed interfere with emphasis. For the other speakers, there may be a mild positive correlation between perception of irritation on an exchange and successful perception of emphasis for the emphasized digit. The topic of the interaction of emphasis and irritation on jaw opening remains to be explored in more detail. Work is currently being done in connection with jaw opening on utterance-initial and final emphasized digits, as well as investigating interspeaker differences.

In terms of articulatory characteristics of emphasis, the data reported here indicate that there is a significant difference between hits and misses, with hits having significantly greater jaw opening than misses. This would suggest that greater jaw opening is linked with the changes in the acoustic speech signal that listeners associate with contrastive emphasis. The finding of a difference in jaw opening between hits and misses does not answer the question why jaw opening did not increase for all the middle digits that were intended by the experimental design to be emphasized, that is, why there were misses in the first place. It is interesting to note that an articulatory characteristic of the type of corrective emphasis elicited in this study is an increased jaw opening on the digit perceived by listeners as emphasized, relative to that of the mean jaw opening of all the digits in the exchange.

The analysis of the data here suggest that there is an effect on jaw opening due to various factors, that is, number of times a digit was corrected, perception of irritation, and perception of emphasis of the produced acoustic signals, plus different strategies among the different speakers to control for these factors. Also, this paper examines a type of

TABLE 10

The mean irritation scores for the hits and misses, with maximum possible score of 20

Speaker	*Hits*	*Misses*
1	1.25 (6%)	5.25 (26%)
2	6.45 (32%)	4.25 (21%)
3	5.25 (26%)	6.29 (32%)
4	7.61 (38%)	3.89 (20%)
Total	6.12 (31%)	4.69 (24%)

irritation associated with the particular task of repeated corrections. Obviously, there are other types of emotion, including other types of irritation, in which a speaker might react by clenching the jaw instead of opening it more presumably resulting in a louder voice. The results of this study suggest that the particular type of irritation elicited is characterized by a more open jaw. The topics of the discrepancy between intended and perceived emphasis, as well as variation among speakers, effect of position of the digit in the utterance, the relationship between jaw opening and other types of emotion, and the relationship between jaw motion and acoustic correlates of emphasis, are currently being investigated by the authors as part of an ongoing research project.

In summary, the demands of nonread dialog-exchange affect speaker's jaw movement patterns for producing word emphasis. It was found that jaw opening increased as the speaker (1) emphasized middle digits, (2) repeated the same correction, and (3) became more irritated. Thus, jaw opening seems to be affected both by emphasis and emotion, that is, irritation. The findings suggest a use of the jaw opening gesture to produce both linguistic or paralinguistic and extralinguistic information, that is, word emphasis and the emotional tenor of the dialog itself.

First received: October 3, 1997, revised manuscript received: April 20, 1998; accepted: May 25, 1998

REFERENCES

BARICK, H. C. (1979). Crosslinguistic study of temporal characteristics of different types of speech material. *Language and Speech*, 20, 116–126.

BECKMAN, M. E. (1995). A typology of spontaneous speech. *Proceedings of ATR International Workshop on Computational Modeling of Prosody for Spontaneous Speech Processing*, pp. 2.23–2.34.

BECKMAN, M. E., & EDWARDS, J. (1992). Intonational categories and the articulatory control of duration. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech perception, production and linguistic structure* (pp. 359–375). Kyoto: Ohmsha.

BECKMAN, M. E., EDWARDS, J., & FLETCHER, J. (1992). Prosodic structure and tempo in a sonority model of articulatory dynamics. In G. Docherty & D. R. Ladd (Eds.), *Papers in Laboratory Phonology II: Segment, gesture, prosody* (pp. 68–86). Cambridge, U.K.: Cambridge University Press.

COHEN, K. B., BECKMAN, M. E., EDWARDS, J., & FOURAKIS, M. (1995). Modeling the articulatory dynamics of two kinds of stress. *Journal of the Acoustical Society of America*, 98, 2894 (A).

Author: ?Editor(s) and place?

COOPER, W. E., EADY, S. J., & MUELLER, P. R. (1985). Acoustic aspects of contrastive stress in question-answer contexts. *Journal of the Acoustical Society of America*, 77, 2142–2155.
EDWARDS, J., BECKMAN, M., & FLETCHER, J. (1991). The articulatory kinematics of final lengthening. *Journal of the Acoustical Society of America*, 89, 369–382.
ERICKSON, D. (to appear). Effects of contrastive emphasis on jaw opening. *Phonetica*.
ERICKSON, D., & FUJIMURA, O. (1996a). Maximum jaw displacement in contrastive emphasis. *Proceedings of ICSLP96*, 1, 141–144 (Philadelphia).
ERICKSON, D., & FUJIMURA, O. (1996b). On defining emphasis. *Paper presented at Fifth Conference on Laboratory Phonology, Northwestern University, Chicago, IL*.
ERICKSON, D., & LEHISTE, I. (1995). Contrastive emphasis in elicited dialog: Durational compensation. *XIIIth International Congress of Phonetic Sciences, Stockholm, Sweden*. 4, pp.352–355.
ERICKSON, D., LENZO, K., & FUJIMURA, O. (1994). Manifestations of contrastive emphasis in jaw movement. *Journal of the Acoustical Society of America*, 95, 2822 (A).
FOWLER, C. A. (1983). Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in monosyllabic stress feet. *Journal of Experimental Psychology: General*, 112, 386–412.
FOX, R., & JOSEPHSON, J. (1992). An abductive articulatory recognition system. *Laboratory for Artificial Intelligence Research Technical Report. The Ohio State University*.
FOX, R., JOSEPHSON, J., & ERICKSON, D. (1991). An abductive articulatory system: A preliminary study. *Journal of the Acoustical Society of America*, 90, 2311 (A).
FUJIMURA, O. (1990). Methods and goals of speech production research. *Language and Speech*, 33, 195–258.
FUJIMURA, O. (1992). Phonology and phonetics: A syllable based model of articulatory organization. *Journal of the Acoustical Society of Japan*, 13, 39–48.
FUJIMURA, O. (1994). C/D model: A computational model of phonetic implementation. In E. S. Ristad (Ed.), *Language Computations* (pp.1–20). Providence, RI: American Math Society.
FUJIMURA, O. (in press). Neuromuscular simulation and linguistic control. *Bulletin de Communication Parlée*.
FUJIMURA, O., & ERICKSON, D. (1996). Prosodic organization of speech signals: Recent development in the C/D model. *Proceedings of Sigphon, Santa Cruz, CA*.
FUJIMURA, O., ISHIDA, H., & KIRITANI, S. (1973). Computer-controlled radiography for observation of movements of articulatory and other human organs. *Computational Biology and Medicine*, 3, 371–384.
GOLDMAN-EISLER, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. London: Academic Press.
HARRINGTON, J., FLETCHER, J., & BECKMAN, M. E. (in press). Manner and place conflicts in the articulation of accent in Australian English. In M. Broe & J. Pierrehumbert (Eds.), *Papers in Laboratory Phonology 5*.
HARRINGTON, J., PALETHORPE, S., FLETCHER, J., & BECKMAN, M. E. (1996). Competing hypotheses concerning the articulation of stress in English. *Journal of the Acoustical Society of America*, 99, 2494 (A).
JONG, K. de, (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *Journal of the Acoustical Society of America*, 97, 491–504.
JONG, K. de, BECKMAN, M. E., & EDWARDS, J. (1993). The interplay between prosodic structure and coarticulation. *Language and Speech*, 36, 197–212.
KIRITANI, S., ITOH, K., & FUJIMURA, O. (1975). Tongue-pellet tracking by a computer-controlled x-ray microbeam system. *Journal of the Acoustical Society of America*, 57, 1516–1520.
LEHISTE, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.
LEINONEN, L., HILTUNEN, T., LINNANKOSKI, I., & LAAKSO, M-L. (1997). Expression of emotional-motivational connotations with a one-word utterance. *Journal of the Acoustical Society of America*, 102, 1853–1863.

LINDBLOM, B. (1990). Explaining phonetic variation: A sketch of the HandH theory. In H. J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modeling* (pp. 403–440). Dordrecht: Kluwer.
MACCHI, M. (1985). *Segmental and suprasegmental features and lip and jaw articulations*. Unpublished doctoral dissertation, New York University.
MACNEILAGE, P. F. (in press). The frame/content theory of evolution of speech production. *Brain and Behavioral Sciences*.
NADLER, R., ABBS, J. H., & FUJIMURA, O. (1987). Speech movement research using the new x-ray microbeam system. *Proceedings of the XIth International Congress of Phonetic Sciences, 1* (pp.221–224). Tallin, Estonia, USSR.
OSHIMA, K., & GRACCO, V. L. (1993). Jaw contribution to stress production. In Haskins Laboratories (Ed.), *3rd Seminar on Speech Production: Models and Data*. Old Saybrook, CT.
PATEL, A. D., LÖFQVIST, A., & NAITO, W. (in press). The acoustics and kinematics of regularly-timed speech: Testing an information-based theory of speech timing. *Journal of the Acoustical Society of America*.
SCHERER, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, 99, 143–165.
SCHULMAN, R. (1989). Articulatory dynamics of loud and normal speech. *Journal of the Acoustical Society of America*, 85, 295–312.
SPRING, C., ERICKSON, D., & CALL, T. (1992). Emotional modalities and intonation in spoken language. *Proceedings ICSLP92*, pp. 679–682.
SUMMERS, W. V. (1987). Effects of stress and final consonant voicing on vowel production: Articulatory and acoustic analyses. *Journal of the Acoustical Society of America*, 82, 847–863.
TULLER, B., & FOWLER, C. (1980). Some articulatory correlates of perceptual isochrony. *Perception and Psychophysics*, 27, 277–283.
VATIKIOTIS-BATESON, E., & KELSO, J. A. S. (1992). Rhythm type and articulatory dynamics in English, French, and Japanese. *ATR Technical Report TR-A-0147*. ATR Auditory and Visual Perception Research Laboratories, Kyoto, Japan.
WESTBURY, J. (1991). The significance and measurement of head position during speech production experiments using the x-ray microbeam system. *Journal of the Acoustical Society of America*, 89, 1782–1791.
WESTBURY, J. (1994). *X-ray microbeam speech production database user's handbook*. Waisman Center on Mental Retardation and Human Development, University of Wisconsin, Madison, WI.
WESTBURY, J., & FUJIMURA, O. (1989). An articulatory characterization of contrastive emphasis. *Journal of the Acoustical Society of America*, 85 (Suppl.1), S98.
WESTBURY, J., MILENKOVIC, P., WEISMER, G., & KENT, R. (1990). X-ray microbeam speech production database. *Journal of the Acoustical Society of America*, 80, 3SP17.
WILLIAMS, C. E., & STEVENS, K. N. (1972). Emotions and speech: Some acoustic correlates. *Journal of the Acoustical Society of America*, 52, 1238–1250.

Author: ?Editor(s)?

Author: This does not match bottom of p.3

Author: ?Editor(s) and place?