3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

MANIFESTATIONS OF CONTRASTIVE EMPHASIS IN JAW MOVEMENT IN DIALOGUE

Donna Erickson[a] *Kevin Lenzo*[b] *& Masashi Sawada*[c]

[a]Speech & Hearing Science, The Ohio State University, Columbus, OH 43210-1002 USA
[b]ATR Interpretative Telecommunication Research Laboratories, Kyoto 619-02, Japan
[c]Center for Cognitive Science, The Ohio State University, Columbus, Ohio 43210, USA

ABSTRACT

Using an automatic tool for measuring and segmenting movement segments of x-ray microbeam pellet tracings, we measured various dimensions of jaw movement in the vertical direction during productions of contrastive emphasis in elicited dialogue situations. We find that the measures maximum vertical displacement and total area seem to correlate highly with contrastive emphasis, and are sufficient for differentiating between syllables spoken with contrastive emphasis and those spoken without contrastive emphasis. Moreover, these measures increase as the speaker is asked to repeat the same correction two to five times.

I. INTRODUCTION

In a previous study [1,2], we measured various aspects of jaw movement during productions of contrastive emphasis in "read" speech, using the MicroTracker (μTracker) analysis tool. This tool automatically segments and measures movement segments of x-ray microbeam pellet tracings. Movement segments with respect to mandible height (perpendicular to the occlusal plane) are defined to be delimited by contiguous time-function maxima or minima. The μTracker measures movement segments in terms of several position, excursion distance, velocity, duration, and curve area measurements. The measurements currently encompass three time domains within the movement segment: the segment as a whole, the initial portion from the left boundary to the maximum, and the final portion from the maximum to the right boundary. Currently there are 28 measures. Figure 1 shows six of these. Total area and maximum vertical displacement (Fig. 1a) are measures pertaining to the syllable as a whole, while initial and final excursion, and initial and final half horizontal area (Fig. 1b and 1c) are measures pertaining to the initial and final portions of the syllable. Maximum vertical displacement (Fig. 1a) is measured from a theoretical closure of the jaw[1] to the maximum opening during the vocalic gesture. The measures initial excursion and final excursion (Fig. 1b) are the relative displacements of jaw opening as measured from the onset and offset positions respectively of jaw movement to the maximum displacement. Initial half horizontal area (Fig. 1b) is the area of the initial portion of the movement segment as measured from a straight horizontal line connecting the beginning edge to the point in time of maximum displacement, and final half horizontal area is the area of the final portion as measured from a straight horizontal line placed at the lower of the two edge positions (Fig. 1b and 1c.). (Final half horizontal area is defined as shown in Fig. 1b or 1c).

In our study of "read" speech, we reported that the measures, final excursion and final half horizontal area, (measurements relating to the final portion of the syllable), showed a consistently higher correlation with contrastive emphasis than the same measurements relating to the initial portion of the syllable, or to the syllable as a whole.

In this study with "elicited" dialogue, we ask if speakers use similar patterns of jaw movement to produce contrastive emphasis.

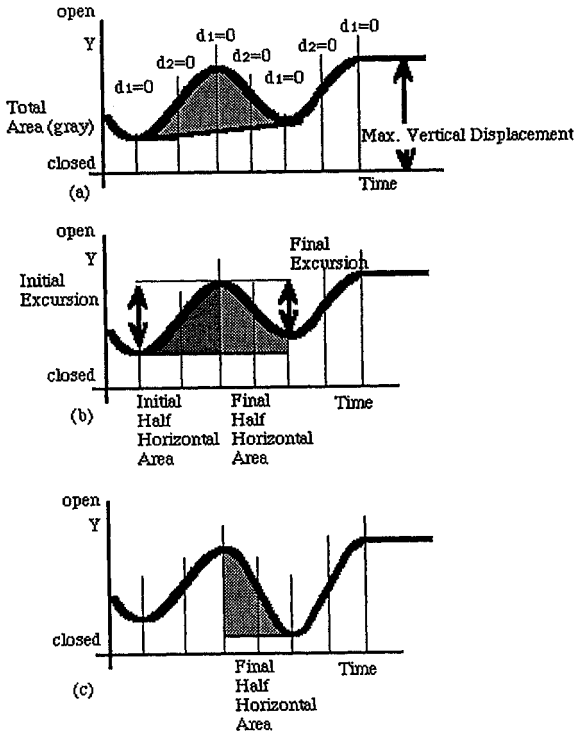

Fig. 1. Schematization of segmentation of jaw motion according to first and second time derivatives is shown in 1a. The μTracker displays jaw opening as peaks, rather than valleys; the 0-point of the y-axis represents closed jaw, with the upper part of the y-axis representing opened jaw. Movement segment dimensions are shown in 1a-c: (a) total area & maximum value, (b) initial & final excursion; initial & final half horizontal area, and (c) final half horizontal area, when the jaw movement of the final portion of the movement segment is lower than that of the onset of the initial portion of the movement segment.

II. METHODS

We made simultaneous acoustic and articulatory recordings (using the x-ray microbeam facilities at the University of Wisconsin [3]) of four American English speakers engaged in dialogue involving target phrases of the types, "599 Pine [Street]". We attempted to elicit contrastive emphasis on each of the initial, medial, and final digits, and also to elicit the phrase without contrastive emphasis on any of the digits. In this paper, we report on the jaw movement patterns of one of these speakers.[2] The speaker produced 79 target phrases in all.

[1] The "zero-point" refers to jaw occlusion. This is always a fixed position for each speaker.

[2] Some preliminary articulatory and acoustic findings were reported for a different speaker in [4].

The first author, who elicited the responses, instructed the speaker to pretend that this was a telephone conversation, and to reply to the questions by reading the prompt on the monitor; she asked the speaker to "not read but to try to get the correct information across" if she (the elicitor) was having problems hearing the response clearly. The elicitor sat out of sight but within hearing distance of the speaker. For the subset of the data involving a variation of the target phrase-type, "5 9 9 Pine [Street]", the elicitor misunderstood the speaker's response by repeating the digit sequence with the initial, medial, or final digit incorrect. The speaker responded by giving the correct information, without reading the monitor prompt. Sometimes the elicitor asked the speaker to repeat the correct digit sequence five or six times.

We assumed that the speaker would first produce the target phrases with no contrastive emphasis, and then in response to the elicitor's misunderstanding of the phrase, with emphasis on one of the digits.

In the previous study with read speech, we used a similar corpus, but the speaker had to utter both the question and the answer in sequence[3]. The prompt for the read speech came with capital letters on the digit to be emphasized. Informal perception tests, as well as our subjective impressions, indicated that the placement of contrastive emphasis could be perceived as that intended by the experimental paradigm, and that the "yes statements" (without contrastive emphasis) were also perceived as having no contrastive emphasis.

For the elicited dialogue data of the one speaker we analyzed in this study, we ran a perception test with five listeners, asking them to circle the digit (or "pine") which they heard as being corrected in the phrase "5 9 5 Pine [Street]", and to circle a question mark if they were not sure they heard any correction. Of the 64 target phrases that were produced in response to the elicitor's request for correction, only 29 were heard by at least 4 of the 5 listeners as having contrastive emphasis on a digit; of the 15 target phrases that were produced in response to the elicitor's first question, i.e., were read from the monitor-displayed prompt and were intended to be spoken without contrastive emphasis, only 10 were heard by at least 4 of the 5 listeners as having no contrastive emphasis. Listeners heard the largest number of instances of contrastive emphasis for those target phrases which were intended by the experimental paradigm to have the middle digit corrected.

In the present study, we measured the duration, excursion, position and velocity of jaw movement in the y-direction of the target phrases, using the μTracker tool. We show a sample display of the μTracker output (Figure 2) for the speaker reading the phrase, "I work at 5 9 5 Pine Street", in response to the elicitor's first question.

The pattern of jaw movement for this speaker's initial responses to the elicitor, as read from a monitor prompt, is very similar to that of the three speakers reported in our earlier study with read speech. We see regular movement segments, which have a monotonic upward curve followed by a monotonic downward curve, the peak corresponding to vowel articulation for the syllable nucleus.

This pattern of regular movement segments seems to change as the speaker is forced to repeat his statements, often expressing more emotion. The phrase in Fig. 3 is in response to the elicitor's fourth request for clarification. We see irregular movement segments (segments 6, 7, 9 & 11) interspersed with the regular (more like sinusoidal) movement segments (segments 5, 8, 10, & 12). When we listen to the phrase, we hear definite phrase boundaries between the first "5" and the emphasized "NINE", and also between the last "5" and "Pine". We think these irregular movement segments carry important prosodic information about phrase boundaries and contrastive emphasis. For purposes of this analysis, however, we measured the simple monotonic parts only. For future work, we intend to analyze these irregular movement segments within the theoretical construct of the C/D model [6].

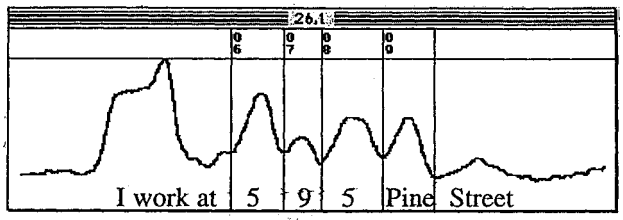

Fig.2. Sample μTracker output of target reference phrase, showing movement segments measured. Jaw excursions are expressed as magnitudes, shown as peaks, rather than valleys. Time is displayed along the x-axis, and jaw excursion (in arbitrary units) is displayed along the y-axis.

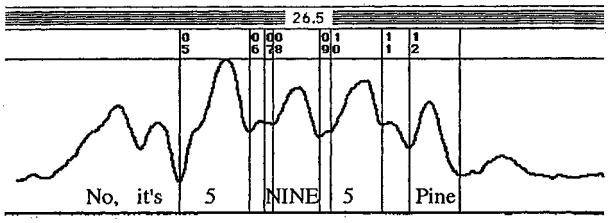

Fig.3. Sample μTracker output of target phrase with emphasis heard by listeners on the middle digit. This is the fourth time the speaker has repeated this target phrase.

For the subset of data for which at least 4 of the 5 listeners heard 7 instances of the phrase "5 N 5 Pine [Street]" with the middle digit corrected and 4 instances of the paired reference phrase "5 9 5" with no contrastive emphasis, we analyze the regular movement segment measurements by correlating the measures with contrastive emphasis. We use the correlation procedure developed for read-speech [1]. We assign a value of "1" to words in the target phrase perceived as having contrastive emphasis and a value of "0" to words in the target phrase not perceived as having contrastive emphasis. We correlate binary values of 1 or 0 (emphasis or no emphasis) with the scalar values of the other 28 measurements of mandible movement.

III. RESULTS AND DISCUSSION

The measures maximum vertical displacement and total area (pertaining to the entire syllable) show the highest correlation with contrastive emphasis for these "5 NINE 5 Pine [Street]" phrases, with correlation coefficients of r=0.72 and r=0.71 respectively. The measures final excursion and final half horizontal area, (which had the highest correlation for the read speech data in the earlier study), have fairly high correlation coefficients of r=0.63 and r=0 .61 respectively. We cite the results of our correlation analysis with contrastive emphasis in elicited dialogue with caution, since the sample size is one speaker, with ten productions of one type of target phrase with middle-digit corrections. (The earlier study with read speech involved three speakers with a total of 223 target phrases, including initial, middle, and final-digit corrections.)

In Fig. 4 we show the two measures, maximum vertical displacement and total area, for comparing the amount of jaw movement involved in producing contrastive emphasis on middle digits as opposed to the amount used in producing middle digits in target phrases without contrastive emphasis. We plot these values for

[3] The original data for this corpus was collected by Westbury & Fujimura [5].

the middle digits of all phrases which were well-perceived to have no contrastive emphasis (n=10) and the middle digits of all target phrases which were well-perceived as emphasized (n=14).

We see a clear grouping between the emphasized digits and the unemphasized digits. It seems the speaker lowers his jaw more on syllables with contrastive emphasis as opposed to syllables without contrastive emphasis.

In this graph we also indicate by different symbols which response it was, i.e., whether the target phrase was spoken as a second, third, fourth or fifth correction response to the elicitor. We see tendencies for groupings of these responses, with the fifth correction having the greatest values for maximum opening and total area. Our subjective impression in listening to these utterances is that the speaker emphasizes more as he is asked to repeat the same phrase over and over again[4]. It seems that as the speaker expresses more emotion, he opens his jaw even more for producing contrastive emphasis.

In summary, contrastive emphasis, (and emotion) have a robust effect on the amount of jaw movement a speaker uses. We intend to continue this method of analysis, looking at all movement segment measurements for this speaker as well as the other three speakers for whom we have elicited dialogue data, as well as exploring the relation between emotion, contrastive emphasis, and jaw movement. We are exploring how articulatory data of this type relates to phonological constructs, and specifically to a theory of phonetic implementation of prosodic information, see e.g., [8].

ACKNOWLEDGMENTS

This work has been supported in part by a research fund from ATR, International, given to O. Fujimura.

REFERENCES

[1] D. Erickson, K. Lenzo, & O. Fujimura, "Manifestations of contrastive emphasis in jaw movement," J. of the Acoust.Soc. of Am., Vol. 95, p. 2822 (A), 1994.

[2] K. Lenzo, K., & D.Erickson, "Measuring articulatory tract motion," J. of the Acoust.Soc. of Am., Vol. 95, p. 2818 (A), 1994.

[3] R. D. Nadler, J.H. Abbs, & O. Fujimura, O. "Speech movement research using the new X-ray Microbeam system," In The 11th International Congress of Phonetic Sciences, Tallinn, Vol.1, pp. 221-224, 1987.

[4] D. Erickson & O. Fujimura, "Acoustic and articulatory correlates of contrastive emphasis in repeated corrections," Proc. 1992 International Conference on Spoken Language Processing, pp. 835-837, 1992.

[5] J. Westbury, J. & O. Fujimura, "An articulatory characterization of contrastive emphasis," J. of the Acoust.Soc. of Am., Vol. 85, (Supplement. 1), S98, 1989.

[6] O. Fujimura, "Syllable timing computation in the CD Model," Proc. 1994 International Conference of Spoken Language Processing, Yokohama, Japan, Sept., 1994.

[7] C. Spring, D. Erickson, & T. Call, "Emotional modalities and intonation in spoken language," Proc. 1992 International Conference on Spoken Language Processing, pp. 679-682, 1992.

[8] K. Lenzo, "From gestures to features: Computer-aided speech analysis and abductive inferential annotation of microbeam data using articulograms," Laboratory for Artificial Intelligence Research Technical Report, The Ohio State University, 1994.

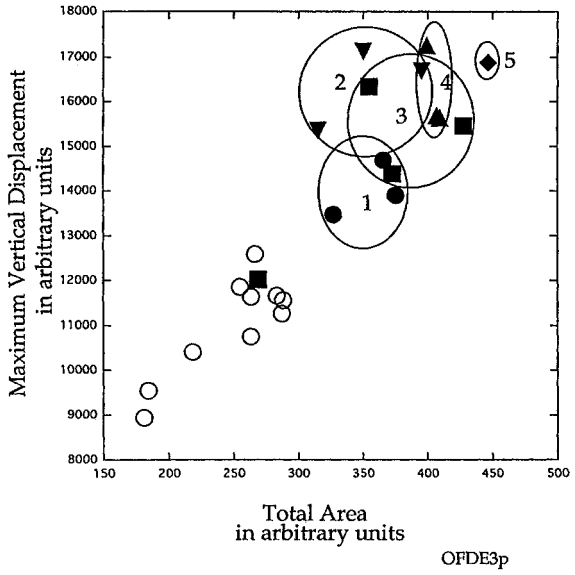

Fig. 4. Comparison of jaw movement on middle digits of well-perceived reference phrases (unfilled circles) and well-perceived medially-emphasized phrases (filled figures). "First corrections" are indicated by filled circles, "second corrections" by filled upside-down triangles, "third corrections" by filled rectangles, "fourth corrections" by filled triangles, and "fifth corrections," by filled diamonds.

[4] In [7] we report on another one of the speakers in this data base who expresses increased emotion as she is asked to repeat the same phrase over and over.