CONTRASTIVE EMPHASIS IN ELICITED DIALOGUE: DURATIONAL COMPENSATION

D. Erickson and I. Lehiste
The Ohio State University, Columbus, Ohio, USA

ABSTRACT

Duration of words in an utterance perceived by listeners as emphasized are longer than the counterpart words in reference utterances (spoken without emphasis); moreover, the non-emphasized words in utterances with emphasis are shorter than the counterpart words in the reference utterances. Thus, emphasis involves temporal rearrangement of all the words in an utterance, not just the word receiving emphasis.

INTRODUCTION

It is generally known that words produced with contrastive emphasis frequently exhibit increases in duration, intensity, and F0 level, e.g., [1]. Here we study the temporal structure of utterances that differ in the presence or absence of contrastive emphasis on one of the words in otherwise identical sentences. The utterances consist of responses involving a 3-digit sequence with the words "five" or "nine" followed by "Pine Street". The utterances were elicited in a dialogue format designed to have the speaker repeat the same correction up to five or six times [2].

Our hypothesis is that emphasis is a phrase level phenomenon; the emphasized digit will be longer in duration relative to the other digits in the sequence; it will also be longer than the corresponding digit in the utterance spoken with no emphasis (hereafter referred to as the "reference utterance" as opposed to the "emphasis utterance"). Since the duration of words in English varies according to their position in the utterance (i.e., phrase final words are longer in duration than the other words), we hypothesize that the percentage increase needed for a word to be perceived as emphasized will also vary.

METHODS

Approximately 38 to 70 target utterances were elicited from each of four speakers of American English in an experimental paradigm that called for contrastive emphasis on one of three digits, and approximately 12 to 18 target utterances intended to be produced as reference utterances. The target utterances were of the type "595 Pine Street", "559 Pine Street", and "959 Pine Street."

The speakers were instructed by the first author to pretend that this was a telephone conversation and to reply to the questions by reading the prompt on the monitor. If the elicitor indicated she was having problems hearing the response clearly, the speakers were asked to "not read the prompt in the monitor screen but to try to get the correct information across according to what the monitor specified." The elicitor sat out of sight but within hearing distance of the speaker. For a subset of responses, the elicitor deliberately misunderstood the speaker's answer 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 more times. We refer to the series of exchanges between elicitor and speaker as a "dialogue set"; it always included one reference utterance and several repetitions of the utterance with the corrected information. A typical dialogue with one speaker is given below. The answer by the speaker to the first question is referred to as the "reference utterance" and is indicated in italics below.

Dialogue 2 (S4)

1.DE: Where do you live?
S4: *I live at 595 Pine Street.*
2.DE: I'm sorry, that was 599 Pine Street?
S4: No, 595 Pine Street.
3. DE: I'm still not getting it. 599 Pine Street?
S4: I live at 595 Pine Street.
4.DE: You're saying, 599 Pine Street?
S4: No, 595.
5. DE: 599 Pine Street, right?
S4: No, 595 Pine Street.

It was assumed 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, with emphasis on one of the digits. We found, however, that it was not always obvious which was the emphasized digit. We ran formal perception tests with 20 listeners and 2 randomizations of the target utterances. The target utterances were the three-digit phrase plus "pine street" which had been extracted from the speaker's response. (Occasionally the speaker would not say "pine street", only the three digit sequence; in which case, only the three-digit sequence was used.) A separate test was made for each of the four speakers in the data base. The listeners' task was to indicate which digit the speaker was making a correction on, and to guess if they were not sure. The results of the perception test indicate that not all of the utterances intended to contain an emphasized digit were identified as such by listeners: only 46% to 70% of possible instances were heard by listeners as carrying contrastive emphasis. We also found that although generally the 3-digit sequence was spoken as part of a single phrase, some of the 3-digit sequences were spoken with each digit as a separate phrase.

Using the Waves+ software, the acoustic signal was digitized and the durations of initial, medial and final digits in the target utterances were measured. In this analysis, we excluded those utterances that were spoken without "Pine Street" or that had phrase breaks between the digits. (Because S1 tended to produce the 3-digit sequence without "Pine Street", to insert phrase breaks between the digits, and generally to produce utterances that were not well perceived by listeners as having emphasis on the intended digit, her data are not analyzed here.)

RESULTS AND DISCUSSION

The durations of "5" and "9" were measured from the reference utterances of the three speakers. These durations did not vary as much as a function of their identity as they did as a function of the position of the digit in the utterance; thus, we averaged together the initial "5's" and "9's", the middle "5's" and "9's", and the final "5's" and "9's".

We measured by position in the phrase the durations of the digits in the reference utterance and those in that particular utterance within the dialogue set that were best perceived by listeners to have emphasis on the intended digit. Figure 1 shows results for one of the speakers. The emphasized digit is always longer in duration than the other digits in the 3-digit sequence, which observation is compatible with findings from other studies of acoustic duration.

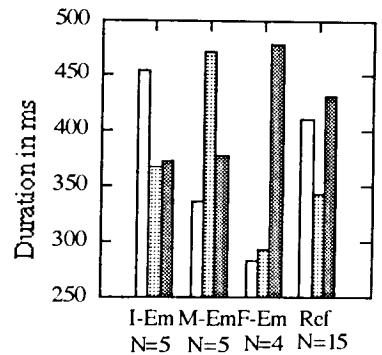

Figure 1. Durations of initial (white), middle (gray), and final (dark) digits for the utterances when the listeners reached the highest agreement on the digit intended to be emphasized in initial, middle and final position, and for the reference utterances.

Moreover, the emphasized digit is longer in duration than the corresponding digit in the utterance spoken with no emphasis. The unemphasized digits are shorter in duration than their counterparts in the reference utterances. For instance, the duration of the final digit in the utterances with final emphasis is clearly longer than the final digit in the reference phrase, and the durations of the initial and middle digits in utterances with the final digit emphasized are decidedly shorter than the initial and middle digits in the reference phrase. This same pattern is seen for the other two speakers (not shown here.)

In order to compare the distribution of durations among the 3 digits across the

different speakers, the durations were calculated in terms of percentages of the total duration of the 3-digit sequence.

Figure 2 shows the results for speaker 3. Note there is a progression in amount of duration needed for a digit to be perceived as emphasized as a function of the position of the digit in the sentence, with the emphasized initial digits constituting 38% of the total duration, middle digits, 40% of the total duration, and final digits, 45% of the total duration.

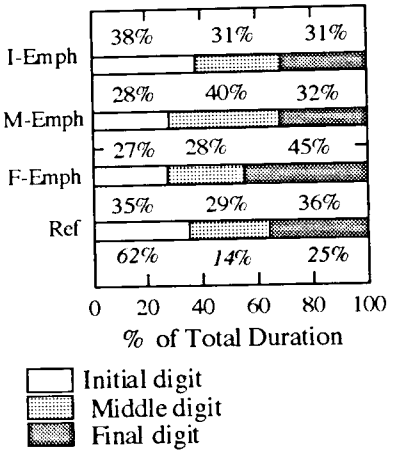

Figure 2. Percent total duration of digits displayed in Figure 1. Percentages are given above each horizontal row; the percentages in italics below the reference bar graph indicate how listeners judged emphasis on the reference utterances.

The bottom row in the graph shows the reference utterances with a break-down of 35%-29%-36% on the initial, middle and final digit respectively. The numbers in italics below the reference utterance indicate how listeners judged emphasis on the reference utterances, given a forced choice task. 62% of the time, listeners assigned emphasis to the initial digit (even though the final digit was longer in duration); 14% of the time, they assigned emphasis to the middle digit, and 25% of the time, to the final digit. It is curious that the duration of the initial digit constitutes 35% of the total duration, yet 62% of the time was heard by listeners as emphasized. Other acoustic cues of intensity and F0 also influence the perception of emphasis. Measurements of peak intensity (in rms) for each digit for this speaker show relatively greater intensity on the initial digit than on the other digits of the reference utterance; this must also be affecting the listeners perception of emphasis. For the other two speakers also, listeners consistently assigned emphasis to the initial digit , even though the initial digit made up approximately only one third of the total duration. Intensity and F0 measurements remain to be made for these speakers.

Table 1 compares the percent of total duration of the 3-digit sequences in the emphasized utterances with those in the reference utterances for each of the three speakers. All three speakers show strikingly similar patterns of duration: the emphasized digit is always greater than its unemphasized counterpart in the reference utterance, and the other digits in the emphasized utterance are always shorter than their counterparts in the reference utterance. The one exception to this is the duration of the middle digit of the utterance with the initial digit emphasized (for speaker 3), which is slightly larger (2%) than the middle digit in the reference utterance.

There is a tendency, especially for speakers 2 and 3, for the initial digit to require less of an increase in duration compared to the middle or final digits in order for it to be heard as an emphasized digit. We wondered why this might be. It may be that the initial emphasized digit increased in duration only by 1% - 3% (for speakers 2 and 3) because when the initial digit made up approximately one third of the total duration of the reference utterance, it was heard as emphasized by over 50% already of the listeners. Thus, only a slight increase in duration would be needed for the initial digit to be heard as emphasized.

Also, it seems that, at least for speakers 2 and 3, a greater increase in duration is required for the middle or final digit to be heard as emphasized than the initial digit.

In summary, it seems that emphasis involves rearrangement of the durational relationships within the utterance, not just an increase in duration of the emphasized item. The emphasized item is increased, and duration is taken off from the other items; the amount of increase/ decrease varies according to the position of the word in the phrase. We suggest that by lengthening the emphasized word and shortening the other words within the utterance, the speaker maximally differentiates the utterance, thus increasing the chances that emphasis will be perceived by listeners.

Table 1. Comparison of the percentage of a digit constituted of the total duration of the 3-digit sequence in reference utterances and utterances with an emphasized digit. Data are shown for speakers S2, S3, and S4.

S2	Initial	Middle	Final	Initial	Middle	Final
Reference (N=8)	34%	32%	34%			
Initial Emphasis (N=1)	35%	31%	34%	**+1%**	-1%	0%
Middle Emphasis (N=3)	29%	41%	30%	-5%	**+9%**	-4%
Final Emphasis (N=4)	27%	28%	45%	-7%	-4%	**+11%**

S3	Initial	Middle	Final	Initial	Middle	Final
Reference (N=15)	35%	29%	36%			
Initial Emphasis (N=5)	38%	31%	31%	**+3%**	+ 2%	- 5%
Middle Emphasis (N=5)	28%	40%	32%	-7%	**+11%**	-4%
Final Emphasis (N=4)	27%	28%	45%	-8%	-1%	**+9%**

S4	Initial	Middle	Final	Initial	Middle	Final
Reference (N=17)	33%	32%	35%			
Initial Emphasis (N=8)	39%	30%	31%	**+6%**	- 2%	- 4%
Middle Emphasis (N=5)	32%	37%	31%	-1%	**+5%**	-4%
Final Emphasis (N=4)	30%	29%	41%	-3%	-3%	**+6%**

References

[1] Lehiste, I.(1970). Suprasegmentals. Cambridge, MA: MIT Press.

[2]Erickson, D., Lenzo, K., & Sawada, M. (1994). Manifestations of contrastive emphasis in jaw movement in dialogue. Proc. International Conference of Spoken Language Processing, Yokohama, Sept. 1994.

Acknowledgments

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