Contrastive Emphasis: Comparison of Pitch Accents with Syllable Magnitudes

C. Menezes[1,3], *D. Erickson*[2] *& O.* Fujimura[1]

[1]Department of Speech and Hearing Sciences
Ohio State University, USA
[2]Gifu City Women's College, Japan
[3]NTT Communication Science Laboratory,
NTT Corp., Japan
{menezes.5; fujimura.1}@osu.edu; erickson@gifu-cwc.ac.jp

Abstract

Contrastive emphasis as elicited from a semi-spontaneous dialogue paradigm was studied with respect to pitch accents and syllable magnitude. In this study, syllable magnitude is defined within the C/D model framework whereby the magnitude of the syllable is related to the displacement of the jaw from the occlusal plane. Articulatory data were collected using the x-ray microbeam facility at the University of Wisconsin. Fundamental frequency patterns were extracted using WAVES+, and the pitch patterns were transcribed using the ToBI system. The results indicate that for all speakers syllable magnitude increases with emphasis in a linear fashion, such that on the average, words that were well-perceived as emphasized have larger jaw opening than those that were poorly perceived as emphasized. The pitch accent associated with the emphasized word differs both within and across speakers. For well-perceived emphasis, the pitch accents can be H*, !H*, L+H*, L*+H, and L*+!H, and these same pitch accents also occur for the same speaker for moderately or poorly-perceived emphasis situations.

1. Introduction

When linguistic units stand out from their environment they are said to be prominent. According to the ToBI method of describing English intonation patterns [1], different types of pitch accents may be aligned to the prominent lexical item. These include H*, L*, L+H*, and L*+H. In ascribing meaning to pitch accents Pierrehumbert *et al*. [1] associate the H* accent with the introduction of a new topic, while the L* is used when the new information is mutually known to both listener and hearer. The L*+H accent is used by a speaker to convey the feeling of uncertainty or "lack of speaker commitment", while the L+H* is used to contrastively emphasize an item. Thus, both the L*+H and the L+H* are said to be used to make an item salient.

Effects of prominence can also be studied by looking at changes in metrical patterns of an utterance, e.g., "syllable magnitude" patterns. Syllable magnitude as defined in this study is based on the framework of the C/D model [2,3,4,5]. The assumption is that the rhythmic organization of an utterance can be represented phonetically as a linear concatenated series of syllable and boundary pulses. The magnitude of the pulses varies in relation to the abstract prosodic strength of each syllable, and is proportional to syllable duration. Syllable duration is derived from articulatory movements, in particular, jaw opening. The study conducted by Erickson *et al* [6] and Erickson [7] lends support to the concept of relating jaw opening to syllable magnitude. Furthermore, using the same database as in this study it was found that jaw opening increases in corrected utterances; in particular, the emphasized digit has the largest jaw opening value within that utterance [8,9]. Previous findings suggest that acoustic manifestations of increased jaw opening may involve among other things changes in formant frequencies [7].

2. Method

2.1. Data recordings

In order to examine the interaction between pitch accents and syllable magnitude (as represented by jaw opening magnitude) in the production of emphasis in spontaneous speech, we analyzed jaw pellet position from data collected at the X-Ray Microbeam Facilities, the University of Wisconsin. (For a description of the microbeam technique, see Westbury[10,11]). The 2.5-3 mm in diameter gold pellet for recording the jaw position was placed on the mandibular incisor. Vertical jaw position was measured as the distance from the maxillary occlusal plane to the center of the pellet sphere as attached to the mandible incisor.

2.2. Speech samples

Spontaneous dialogues eliciting correction of a digit in a street address were recorded from each of four speakers of Midwestern American English (2 men, 2 women). The target phrase was always a 3 digit sequence: "5 9 5", "9 5 9", or "5 5 9". In collecting the data, the speakers were told to pretend this was a telephone conversation in which the elicitor was not able to hear well, and therefore she may have to ask for clarification. The first response of the subject was always read from a monitor display. The entire dialogue (see sample below) lasted 25 seconds, and was recorded continuously with the microbeam pellet tracking. The digit that was to be corrected occurred in the first, second, or last of the three digits in the sequence, but within one dialogue, the correction elicited was consistently on a particular digit. The experiment elicited 12-18 dialogues from each of the subjects. The experiment was designed to also elicit irritation from the speaker. An example of a dialogue is given below:

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.

2.3. Articulatory analysis

Measurements of the lowest vertical jaw position were made for each of the digits using a MATLAB-based software program (Ubedit) developed by Bryan Pardo [12].

In the data analysis reported here, only the dialogues eliciting correction on the second (i.e., middle) digit of the 3-digit sequence were used. Throughout the analysis, the digits "5" and "9" 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 this study showed no significant difference between the amount of jaw opening for "5" and "9" [13].

2.4. Pitch accent analysis

Using WAVES+ and the ToBI transcription method, 2 phoneticians (the first two authors) marked the pitch accents for each of the utterances. The pitch accents were decided using the traditional method for ToBI transcription—listening to the audio signal at the same time looking at the F0 contour. The difference between the L*+H and L+H* pitch accents was a matter of timing—in both cases, there was a rise from low to high pitch but in the L*+H accent, the peak F0 was toward the end of the syllable, whereas for the L+H*, the peak F0 was in the middle of the syllable.

2.5. Emphasis perception tests

The design of the experiment called for speakers to correct street addresses, which were misunderstood by the experimenter. In order to test whether listeners were able to perceive which digit was corrected, perception tests were conducted. Reported here are perception tests done earlier with the same data, as part of acoustic and articulatory studies by Spring, Erickson, and Call [14] and Erickson and Lehiste [15]. Perception tests were run on the 3-digit sequences plus "Pine Street" (excluding the rest of the utterance) uttered by each of the 4 speakers in separate listening test sessions. Two randomizations of the sequences were presented to 20 university students.

3. Results

For this study we only looked at those dialogues that contained middle digit corrections. Emphasis scores obtained from the previous study by author Erickson were divided into three bins corresponding to the subjective categories of 'Well-perceived', 'Moderately-perceived' and 'Poorly-perceived'. All digits that were perceived for emphasis 80% of the time and greater were grouped in the 'Well-perceived' category, those perceived 79% to 40% were included in the 'Moderately-perceived' and all values below 39% were included in the category of 'Poorly-perceived'. Analyses were conducted separately for each speaker.

3.1. Emphasis and syllable magnitude

Fig. 1 shows bar graphs plotted for averaged syllable magnitude (jaw opening) on the vertical axis and perception of emphasis (binned) on the abscissa. The types of pitch accents used by the speaker are depicted within each bar in terms of percentages of occurrences. In this figure it is evident that those digits that were 'Well-perceived' as emphasized had the largest average jaw opening while those digits that were perceived to be less emphasized had less average jaw opening. This was true for all speakers. The correlation of syllable magnitude with perception of emphasis was significant for all speakers (Speaker 1: $r = .830$, $p = .001$; Speaker 2: $r = .707$, $p < .001$; Speaker 3: $r = .676$, $p < .001$ and Speaker 4: $r = .705$, $p < .001$).

3.2. Emphasis and pitch accents

Much variability was seen in the assignment of pitch accents by speakers. Table 1 gives the percentage occurrence of pitch accents for each speaker for the 'Well-perceived', 'Moderately-perceived' and 'Poorly-perceived' emphasis conditions for the middle digit (which was intended by the dialogue protocol to be emphasized). Also in Fig. 1 it can be seen that across the different emphasis situations some speakers do not vary much in the assignment of pitch accents.

4. Discussion

From this study we can see that for all speakers, syllable magnitude as a measure of jaw opening increases with increasing emphasis. At the same time we find that speakers vary with the type of pitch accent they assign to the 'Well-perceived' contrastively emphasized digit. Some speakers choose to use high F0 on the stressed syllable as in H*, or L+H*, where the rise in F0 occurs before the onset of the vowel. On the other hand, some speakers assign low F0 to the stressed syllable as in the L*+H pitch accent, where the rise in F0 begins in the vowel and the peak manifests in the final consonant or later. Thus, there is a difference in strategy as to which aspect of the F0 contour associates with the stressed syllable. Moreover, certain speakers tend to use predominantly one pitch accent for well-perceived emphasis Speaker 2 tends to use L+H* and Speaker 4, L*+H. However, the other 2 speakers use 3 types of pitch accents (almost equally frequently) for digits that were 'Well-perceived' as emphasized. In addition, for all speakers, the pitch accents used for 'Well-perceived' emphasis are also used with 'Moderately' or 'Poorly-perceived' emphasis.

5. Conclusions

American English speakers in producing contrastive emphasis make the word to be emphasized stand out from its environment, presumably so listeners can perceive an increased prominence or salience. Speakers are able to control both the metrical structure (changes in syllable magnitude) and the tonal structure (changes in pitch accent) to make a specific word stand out from the others around it. The interesting finding in this study is that on the average all speakers use increased jaw opening to signal an emphasized word. But the

choice of pitch accents seems to be a matter of individual choice. Furthermore, either a lowering or raising of pitch can be used to indicate emphasis. It could be that in semi-spontaneous elicitations a variety of pitch accents are used to convey contrastive emphasis. Or perhaps the difficulty of the particular task, i.e., to keep repeating the same correction, contributed to the speaker using a variety of pitch accent in order to try to communicate more clearly the correction to the listener. These findings suggest that for American English, which is said to be a syllable-stressed timed language, that metrical structure, as indicated by changes in syllable magnitude patterns, may constitute the underlying prosodic structure of American English, and tonal structure is added on as a matter of individual speaker choice. To what extent this can be said to apply to other syllable stressed timed languages is an interesting question to pursue.

6. Acknowledgements

The authors would like to thank Julie McGory for her advice on the ToBI transcriptions. This work was supported by NTT, Japan.

7. References

[1] Pierrehumbert, J.; Hirschberg, J., 1990. The meaning of intonation contours in the interpretation of discourse. In *Intentions in Communication and Discourse (SDF Benchmark Series in Computational Linguistics*, P.R. Cohen, J. Morgan and M.E. Pollack (eds). MIT Press, 271-311.

[2] Fujimura, O., 1992. Phonology and phonetics-A syllable-based model of articulatory organization. *J. Acoust. Soc. Jpn. (E)* 13, 39-48.

[3] Fujimura, O., 1994, C/D model: A computational model of phonetic implementation. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* (Am. Math. Soc.) 17, 1-20.

[4] Fujimura, O., 2000a. The C/D model and prosodic control of articulatory behavior. *Phonetica* 57, 128-38.

[5] Fujimura, O., 2000b. The C/D model prediction of CVC segmental duration for varied syllable prominence. *Phonetician* 82, 9-21.

[6] Erickson, D.; Fujimura, O.; Dang, J., 1999. Articulatory and acoustic characteristics of emphasized and unemphasized vowels. *J. Acoust. Soc. Am*., 106, 2241.

[7] Erickson, D. to appear. Articulation of extreme formant patterns for emphasized vowels. Phonetica.

[8] Menezes, C.; Pardo, B.; Erickson, D.; Fujimura, O., accepted. Changes in syllable magnitude and timing due to repeated correction.

[9] Mitchell, C.J.; Menezes, C.; Williams, J.C.; Pardo, B.; Erickson, D.; and Fujimura, O., 2000. Changes in syllable and boundary strengths due to irritation. ISCA Workshop on Speech and Emotion. Newcastle, 98-101.

[10] Westbury, J.R., 1994. *X-ray microbeam speech production database user's handbook*. Waisman Center on Mental Retardation and Human Development. University of Wisconsin, Madison.

[11] Westbury, J.R.; Fujimura, O., 1989. An articulatory characterization of contrastive emphasis. *J. Acoust. Soc. Amer*. 85 (S1), s98 (A).

[12] Erickson, D.; Fujimura, O.; Pardo, B., 1998. Articulatory correlates of prosodic control: Emotion and emphasis. *Language and Speech* 41, 395-413.

[13] Erickson, D., 1998. Effects of contrastive emphasis on jaw opening. *Phonetica* 55, 147-169.

[14] Spring, C.; Erickson, D.; Call, T., 1992. Emotional modalities and intonation in spoken language. In J. Ohala (ed), *Proc. Internat. Conf. On Spoken Lang. Proc*. Alberta, Canada, 679-682.

[15] Erickson, D.; Lehiste, I., 1995. Contrastive emphasis in elicited dialogue: durational compensation. *Proc. 13th Internat Congress of Phonetic Sci*, Stockholm, v4, 352-355.

Table 1: Percentage of pitch accents for different speakers and different perception of emphasis. In some cases there were no pitch change perceived on the digit (Speaker 2, 29% in poorly-perceived, Speaker 3, 6% in well-perceived and 9% in poorly-perceived).

Speaker	Well-Perceived				Moderately-Perceived					Poorly-Perceived				
-	L*+H	L+H*	H*	!H*	L*+H	L+H*	L*+!H	H*	!H*	L*+H	L*+!H	H*	!H*	L*
1	25	-	25	50	-	-	100	-	-	-	29	-	71	-
2	-	91	9	-	-	50	-	50	-	-	-	57	-	14
3	31	38	25	-	-	-	-	-	100	-	36	-	55	-
4	90	10	-	-	50	-	-	25	25	46	9	18	18	9

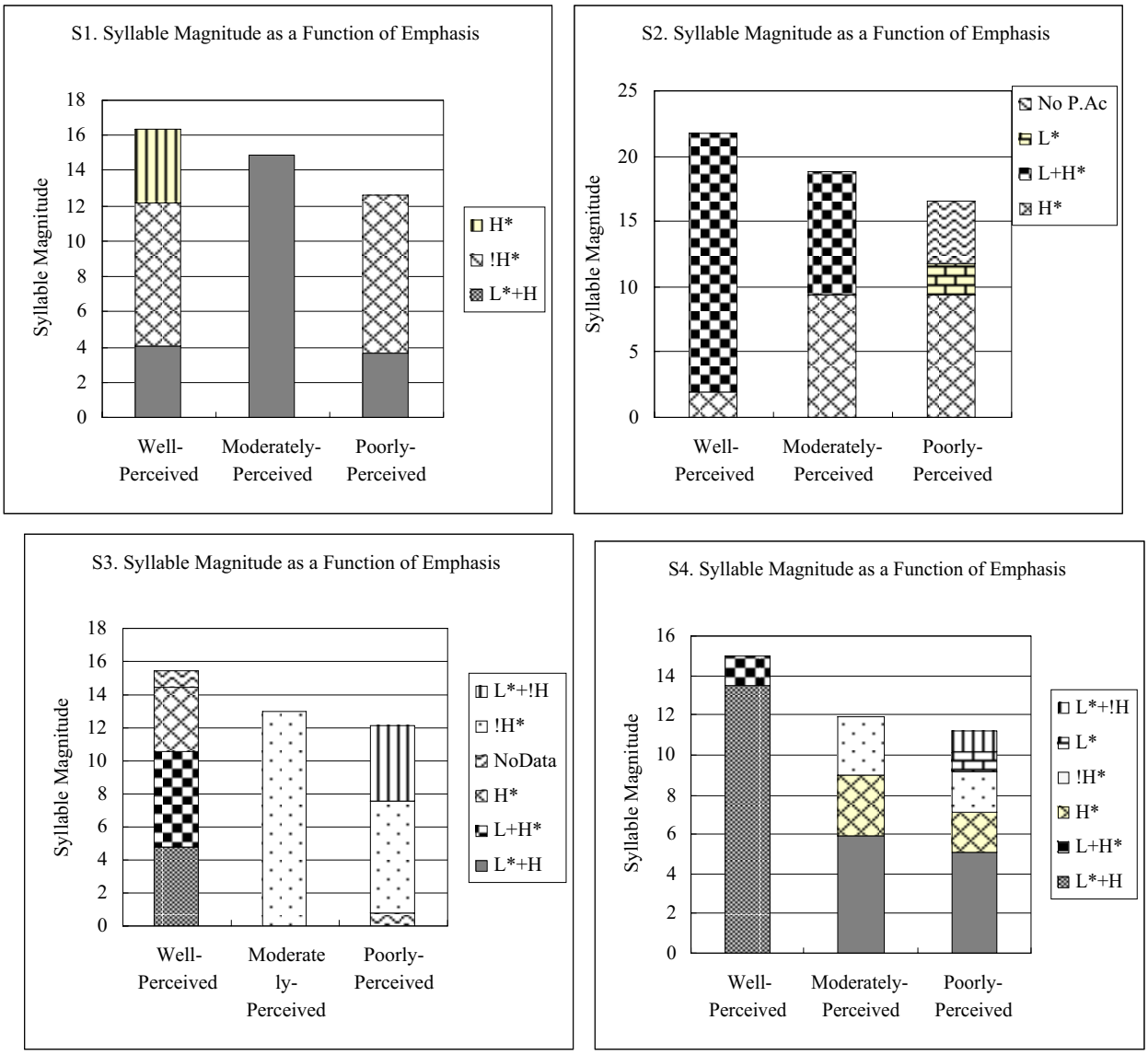

Figure 1: Syllable magnitudes and pitch accents for four speakers (S1, S2, S3, S4) on middle digits (which were intended to be emphasized) as a function of how well emphasis was perceived by listeners. 'No P.Ac' represent cases where no pitch accent occurred.