Comparing the Metrical Structure of a Digit Sequence in American English: Articulatory vs. Acoustic Analysis

Caroline MENEZES*

SUMMARY: This study compares the acoustically and articulatorily derived rhythmic structure of American English utterances of a short three digit phrase uttered within a semi-spontaneous dialogue. The Converter/Distributor model of Osamu Fujimura forms the framework for determining the articulatory rhythm. Within this model, syllable stress is measured as a scalar representation of articulatory strength. Articulatory syllable duration is an abstract representation of syllable magnitude and is directly proportional to it. Results show that the acoustic rhythmic structure varies greatly from the articulatory rhythmic structure for these semi-spontaneous utterances. Emphasis was shown to change syllable durations, syllable magnitudes and boundary magnitudes.

Key words: C/D model, metrical structure, articulatory analysis, syllable magnitude, rhythmic analysis

1. Introduction

According to Hayes (1995), the metrical structure of an utterance is the base on which its phonetic and phonological features are organized, and he claimed that “stress is the linguistic manifestation” of this metrical structure. Earlier, Liberman and Prince tackled the temporal structure of speech as a hierarchy of stress and meter (Liberman 1975, Liberman and Prince 1977, Prince 1983). The metrical patterns of Liberman are basically trees (later metrical grids were also formulated) with binary branching of relatively strong (s) and weak (w) nodes (an element is strong only when presented with another element that is weak). The constituents of the trees are ordered in time, and the trees are rooted, that is, there is only one constituent at the highest level. The main stress is always the most prominent terminal element. In his dissertation Liberman (1975) aligned the highest tone to the strongest syllable with the longest duration. It is important to note here that Liberman and Prince (1977) in their proposal of the Metrical theory deviate clearly from some earlier concepts of linguistic “stress” in two important points. First, stress could act above the word level, in contrast to Bolinger (1958) and Fry (1955), for whom the domain of stress was only lexical. Bolinger also believed that sentence stress was nothing more than a handle for pitch-accents. Second, according to the metrical theory, stress in English has phonetic reality in the rhythmic structure and is independent of the placement of intonation contours. In addition, metrical theory is a relational theory, and as such, the phonetic implementation of prominence cannot be confined to local articulatory gestures or sound quality but rather to larger units such as phrases and utterances.

To study the metrical structure of semi-spontaneous utterances the frame work provided by the Converter/Distributor model (C/D model) proposed by Osamu Fujimura (1992, 1994, 2000a, 2000b) is used. The C/D model is a syllable-based phonetic implementation of the temporal organization of speech utterances. For details of this model, please read the articles presented in this volume by Osamu Fujimura. The C/D model generates both articulatory and acoustic signals as outputs from linguistic form inputs with additional numerical parameter specifications. These numerical specifications account for extra-linguistic features like speaker style, rate of speech, emotional state of the speaker, etc. Like the metrical theory, the C/D model proposes that the rhythmic structure of an utterance forms the organizational framework upon which segmental aspects of the speech signals are superimposed. It assumes that the rhythmic organization of an utterance can be represented by a linear string of syllables and boundaries.

*University of Toledo (トレド大学)
Comparing the Metrical Structure of a Digit Sequence in American English

Pulses of varying magnitudes.

Syllable magnitudes are estimated directly from the maximum vertical displacement of the jaw, such that the greater the deviation of the jaw, the larger the magnitude of the syllable. In other words, mandibular movements represent the base function control of rhythm. Syllable duration is an abstraction of syllable magnitude and is directly proportional to it. According to the theory, the syllable magnitude also affects various segmental duration measures on the acoustic signal within a given syllable.

Boundary magnitudes, on the other hand, are calculated as the duration of gaps between two consecutive syllables. The duration of boundaries also varies continuously and is proportional to the magnitude of the boundary. This one-dimensional pulse train of syllable and boundary pulses is assumed to carry at least for the first approximation, all information about the rhythmic structure of an utterance and is congruent with a phonetic metrical grid.

Tonal specifications and lexical and phrasal accent features are implemented by the melodic base function which is independent of the rhythmic base function. Other melodic aspects of the base function include vowel-to-vowel articulatory movement and switching between voiced and unvoiced phonetic status (Fujimura 2000a, 2000b).

Previous studies have shown that syllable magnitude computed from jaw opening increases with prominence resulting from discourse factors such as contrastive emphasis (Erickson 1998, 2002, Erickson et al. 1998, 1999, Westbury and Fujimura 1989, Menezes et al. 2003, Menezes 2003). Menezes 2003 studying the same data base has shown that syllable magnitude changes with repeated emphasis and perceived irritation. They also noted that magnitudes and timing of boundaries also changed suggesting phonetic phrasal reorganization.

This current study focuses on the phrasal reorganization in semi-spontaneous dialogues where repeated corrections were elicited on the middle unit of a three digit street address. The algorithm for computing the hypothetical syllable magnitude and duration values remains the same as that used in the Mitchell et al. (2000) study. This algorithm allowed us to create syllable triangles from the apex of the syllable pulses such that there was at least one point in the dialogue where edges of two syllable triangles were contiguous but no overlap could occur over the entire dialogue. The base of these triangles formed the abstract syllable duration and gaps between syllable triangles were measured as boundaries. Thus we were able to quantitatively determine the magnitude and occurrences of boundaries. Additionally, timing of each syllable pulse was determined from the velocity measures of the crucial articulators of the onset and coda consonants (See Bonaventura and Fujimura 2007 for “iceberg” algorithm).

2. Method

2.1 Data Recordings

Acoustic and articulatory data reported in this study were collected at the X-Ray Microbeam Facilities at the University of Wisconsin (For a description of the microbeam technique, see Westbury 1994, Fujimura et al. 1973, Kiritani et al. 1975). Gold pellets of approximately 2.5–3 mm in diameter were placed on the mandibular incisor to measure the precise vertical displacement of the lower jaw. Vertical jaw position was measured as the distance from the maxillary occlusal plane to the center of the pellet sphere attached to the mandible incisor. Vertical position of the tongue tip and lower lip were measured to determine the timing of the syllable pulse.

2.2 Speech Samples

The data consisted of semi-spontaneous dialogues between the subject and the experimenter wherein the subject was given a specific street address that was to be the target response to the experimenter’s question. The target phrase was always a 3 digit sequence: “5 9 5”, “9 5 9”, or “5 5 9” followed by “Pine Street” (for details on the experimental paradigm see Menezes et al. 2003). Four speakers of Midwestern American English (2 men, 2 women) participated in this study. In collecting the data, the speakers were told to pretend that 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 computer monitor. The entire dialogue (see sample below) lasted 25 seconds, and was recorded continuously with the microbeam pellet tracking. The clarification always occurred on one of the three digits in the address sequence, but within a dialogue the correction elicited was consistent. 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 Inference of Syllable Magnitude and Timing

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 (Mitchell et al. 2000, Menezes et al. 2003, Menezes 2003). The program measures the position value and timing of the minimum vertical position of the jaw for the three digit phrase sequence along with “Pine Street.” It also measures the key points in the vertical positions for the crucial syllable onset and coda articulators (tongue tip for “nine” and lower lip for “five”) to facilitate semi-automatic data processing (see Menezes et al. 2003 for details).

The magnitude of each syllable was set to be proportional to the maximum jaw opening measured as the distance from the occlusal plane to the minimum mandible position (Fujimura and Williams 1998). UBEDIT automatically constructs a syllable pulse for each target syllable and the timing of the pulse was adjusted to the “iceberg value” of the crucial articulators of the onset and coda consonants for that specific syllable. The program then constructs an isosceles triangle (syllable triangle) from the apex of the pulse to compute the abstract syllable duration (Fujimura and Williams 1998, Mitchell et al. 2000, Menezes et al. 2003). To compute boundary magnitudes the arms of the syllable triangles were allowed to expand outwards till arms of two contiguous syllables met at one point in the dialogue. All resultant gaps between syllables were considered to be boundaries and the duration of these gaps directly determined the magnitude of the boundary.

Careful visual inspection of the acoustic waveform and spectrogram and auditory inspection of target utterances were done to isolate the acoustic edges of the three digit sequence and “Pine Street” using PRAAT. Acoustic durations were then extracted and compared with the articulatorily-derived syllable and boundary durations.

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” (Erickson 1998).

3. Results

This study is focused on the rhythmic reorganization of utterances for dialogues where the middle digit of the three digit sequence was controlled for emphasis. This is because this digit is never emphasized in the reference utterance. The reference utterance is the first exchange where the subject provides information to the examiner’s question about where something exists. All other exchanges in the dialogues will be compared to this reference. Another reason for studying this digit is its position in the utterance. It is sandwiched between two computed boundaries and the effect of emphasis on both syllable and boundary durations can be clearly observed.

Figure 1 plots the mean articulatory syllable duration in milliseconds as computed from the C/D model algorithm. The histograms are presented separately for the 4 different speakers. The bars are clustered by the number of exchanges in all dialogues recorded for a given speaker. Comparing all four speakers it is evident at first glance that all speakers have different dialogue characteristics. However, looking deeper we see similarities in the metrical structure. The initial digit in the reference utterance or the first word in the three word phrase always has the largest magnitude for all speakers. However, in the following exchanges we see a significant increase in the duration of the middle digit following correction. Figure 1 also interestingly shows that corrective emphasis on one word in the phrase has the effect of increasing the duration of all other words in the phrase. This sometimes occurs at the cost of the syllable duration in “Pine” which in the target utterance always occurs as a phrase with “street” (e.g., Five nine five, Pine Street).

Furthermore, we see that corrected utterances following the first exchange are speaker specific. Speakers two and four, for example, steadily increase syllable duration for all words in the target phrase with increasing number of exchanges. But Speakers one and three maintain the same durations over the exchanges. Speaker one demonstrates the ideal metrical pattern of alternating syllables of relative stress. This is not evident, however, for the other speakers. Speaker three,
Comparing the Metrical Structure of a Digit Sequence in American English

Figure 1 Bar graphs showing mean articulatory syllable duration for all exchanges where the middle digit was emphasized, separated by speakers. Error bars indicate mean ±/− 1.0 SE.

In particular, maintains the metrical structure of the reference utterance for all exchanges in the dialogue, i.e., the largest syllable always being the initial digit. The metrical pattern observed in Figure 1 cannot be completely explained without observing boundary placement and magnitudes which will be discussed later.

Figure 2 displays the mean jaw opening values separately for all four speakers clustering again by exchange number for all dialogues recorded for a given speaker. According to the C/D model, jaw opening is indicative of syllable strength or magnitude. Like articulatory syllable duration, we see that the largest jaw opening for the reference utterance is consistently on the initial digit for all speakers. Jaw opening for the middle digit is significantly less when compared to the initial digit. Jaw opening for the final digit differs according to speaker. For Speakers one and four, it is relatively stronger than the middle digit, but for Speakers two and three, it is weaker. However, following correction, as in the first exchange, we see a significant increase in jaw opening, again imitating the pattern seen for articulatory syllable duration.

In fact, jaw opening clearly follows the same pattern as observed by the articulatory syllable duration. This indicates that the algorithm to compute the abstract syllable duration from jaw opening values was successful. This in turn validates the C/D model concept of determining syllable magnitude or strength from the articulatory feature of jaw opening.

Figure 3 plots the mean acoustic syllable duration for all four speakers in a fashion to Figures 1 and 2. The acoustic duration of “Pine” was not calculated and is therefore not shown in this figure. The acoustic syllable durations do not follow the same pattern observed on jaw opening values and the articulatory syllable durations derived from it. However, we do see that for three of the speakers (Speakers one, three and four) the initial digit has the largest duration when compared to the other syllables in the reference utterance. Following correction in the first exchange, the emphasis placed on the middle digit also resulted in an increase in acoustic duration. The alternation of relatively strong and weak syllables is more evident in the acoustically-derived metrical pattern when compared to the articulatorily-derived metrical pattern. This is partly due to the large acoustic duration of the final digit of the phrase. This phenomenon is well known as “phrase final lengthening” (Edwards et al. 1991, Beckman and Edwards 1994). There are no empirical studies that discuss how phrase final lengthening affects the strength of the syllable.

Figure 4 plots the mean articulatory boundary duration that was computed as gaps between the derived articulatory syllable durations. Boundary patterns of
特集「音声科学の最前線：C/D モデルその音声学・音形論への衝撃」

Figure 2 Bargraphs showing mean jaw opening for all exchanges where the middle digit was emphasized, separated by speakers. Error bars indicate mean +/- 1.0 SE.

Figure 3 Bargraphs showing mean acoustic syllable durations for all exchanges where the middle digit was emphasized, separated by speakers. Error bars indicate mean +/- 1.0 SE.
the reference utterance show that the boundary between the initial and middle digit has the smallest magnitude, indicating that the initial and middle digit form a unit separate from the final digit. The largest boundary is observed in most instances between the final digit and “Pine” indicating that “Pine Street” forms a separate phrase from the digit sequence. Therefore, our discussion will not deal with this phrase and will be limited to the patterning of the digit sequence only. Emphasis on the middle digit not only changes the syllable duration pattern, it also changes the boundary durations. In general, the boundary before the emphasized digit becomes more prominent, increasing its magnitude and duration. This in effect changes the “association” of the middle digit. It is no longer associated with the initial digit but forms a phrase with the final digit. Speakers three and four who do not show the alternating stress pattern in the articulatory syllable durations appear to use boundaries as realignment points for the metrical structure. A large boundary inserted right after a strong syllable can cause the next syllable to be strong relative to the syllable that follows it. For example the reference utterance for Speaker three shows the final syllable to be weak following a relatively weak middle digit; however, the boundary between the middle and final digit is large. Similarly, in the following exchanges the large boundary before the emphasized digit can append relative strength to the middle digit followed by the weak final digit. Using this strategy, the speaker does not need to increase the duration of the middle digit even when emphasized in order for it to bear prominence in its phrase.

Figure 5 plots the mean acoustic boundary durations that occur between the initial and middle digits and the middle and final digits. Like the difference between the two computations of syllable durations, the acoustic and articulatory-measured boundary durations show different patterns. However, except for Speaker one, the initial and middle digits still appear to be cohesive units separate from the final digit. Emphasis on the middle digit, however, reveals speaker specific variations. Speakers two and three show changes in the minor boundary such that it increases in magnitude similar to what was observed for the articulatory boundary patterns. However, Speakers one and four don’t show any change in the boundary patterning following emphasis on the middle digit. Moreover, the acoustic boundary durations do not help to disambiguate instances when the metrical structure of the utterance violates the alternation of relative stress.
4. Conclusion

This study has attempted to analyze quantitatively and describe phonetically the rhythmic structure of an utterance produced within a semi-spontaneous dialogue. In general, the results have shown that speakers use various patterns when producing a digit sequence phrase, both in the reference condition and in the implementation of corrective emphasis. However, some phonetic features were consistent from speaker to speaker. For example, the initial digit in the reference utterance was always assigned the largest magnitude in the phrase comprised of a three digit sequence. Fougeron and Keating (1997) and later Fougeron (2001) has discussed this phenomenon which they refer to as "initial strengthening." The first syllable in an utterance is articulated with maximum energy. Speakers differed, however, in how they treated the rest of the phrase. Speakers one and four produced the digit sequence with an alternating pattern of large and small magnitudes. Speakers two and three, on the other hand, produced the utterance with a gradual decrease in syllable magnitude from initial to final digit. As discussed above, it is assumed that these two speakers used boundaries to restart the metrical organization. A strong boundary allows the following syllable to be relative stronger in magnitude, even if its duration is shorter than the syllable in front of it.

Speakers also consistently show an increase in both articulatory and acoustic syllable duration and magnitude when emphasis is placed on that specific syllable. In some cases, emphasis on one syllable resulted in a spreading of emphasis to all syllables within the phrase. The increase in syllable magnitude of the corrected digit was clearly observed in the larger jaw movements and the longer syllable durations. This indicates that the algorithm used by the C/D model to extract abstract duration from jaw deviations was successful in relating scalar magnitude values to abstract duration values.

Acoustic syllable durations, however, did not reflect the articulatory energy used to produce the given vowels, and in several instances, did not match the articulatory durations. The syllable magnitude patterns derived from acoustic durations were highly compromised by phrase final lengthening. Determination of syllable durations from jaw deviations was also successful in separating phrase final effects from articulatory syllable durations.

Emphasis was also shown to change the metrical structure of the utterance by changing the magnitude of both syllables and boundaries. Phrasal restructuring also was found to occur when the middle digit was emphasized. Several interesting trends regarding the hierarchy of phrasing could be observed but was outside the scope of this paper.
Comparing the Metrical Structure of a Digit Sequence in American English

Acknowledgements

I would like to acknowledge Osamu Fujimura the author of the C/D Model for introducing me to articulatory phonetics, the guidance he has provided me and the doors of opportunity and learning he opened for me throughout my scientific learning. I would also like to thank Donna Erickson who collected the Blue Pine Street data at the University of Wisconsin.

Notes

1) The research presented in this paper was previously described in my PhD. Dissertation.

References


(Received Aug. 15, 2015, Accepted Oct. 20, 2015)