Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

ELSEVIER

Speech Communication 40 (2003) 71–85

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Changes in syllable magnitude and timing due to repeated correction

Caroline Menezes [a,*], Bryan Pardo [b], Donna Erickson [c], Osamu Fujimura [a]

[a] *Department of Speech and Hearing Science, The Ohio State University, Columbus, OH 43210-1002, USA*
[b] *The University of Michigan, Ann Arbor, Michigan, MI 48109 USA*
[c] *Gifu City Women's College, Gifu City, 501-0192 Japan*

Abstract

In a semi-spontaneous conversational setting, subjects were made 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 signals were recorded by the University of Wisconsin Microbeam Facility for four speakers of American English. By analyzing jaw movements, syllable magnitude and time values were evaluated, to represent the rhythmic organization of the utterance by a linear string of syllable pulses. Preliminary results suggest that not only does the magnitude of the corrected syllable increase by the correction of a digit, but also, in most cases, there is some systematic increase of syllable magnitude both in the corrected digit and other digits in the same utterance, as the same correction is repeated. Considerable difference among different speakers is observed and discussed in terms of syllable magnitude and timing patterns.

Zusammenfassung

In einer teilweise spontanen experimentellen Unterhaltungssituation wurden Testpersonen dazu veranlaßt, wiederholte gesprochene Korrekturen einer Ziffer innerhalb einer Dreiziffernfolge zu produzieren. Die Ziffernfolgen bestanden aus den englischen Zahlworten "five" und "nine" gefolgt von "Pine Street". Artikulatorische und akustische Signale wurden in der Microbeam-Abteilung der Universität Wisconsin für vier Sprecher des amerikanischen Englisch erfaßt. Durch Analyse von Kieferbewegungen wurden Maße von Silbengröße und Zeitwerte bestimmt, um die rythmische Organisation der ausgesprochenen Produktionen im Rahmen eines lineares Feder-Models zu repräsentieren. Vorläufige Ergebnisse weisen darauf hin, daß nicht nur die Silbengröße der korrigierten Silben zunimmt, sondern auch, daß in den meisten Fällen eine kleine systematische Erhöhung der Silbengröße sowohl in den korrigierten als auch in den anderen Silben zu beobachten ist, wenn dieselbe Korrektur wiederholt wird. Merkliche Unterschiede zwischen verschiedenen Sprechern wurden beobachtet und werden hinsichtlich Silbengröße und zeitlichen Mustern diskutiert.

Résumé

Pendant un dialogue semi-spontané, les sujets étaient obligés de répéter la même correction d'un chiffre dans une série de trois chiffres qui se compose de "five" ou "nine" suivi de "Pine Street". Les signaux articulatoires et acoustiques de quatre parleurs de l'anglais américain général furent enregistrés par la X-ray Microbeam (la radiographie au

* Corresponding author.

0167-6393/02/$ - see front matter
PII: S0167-6393(02)00076-6

micro-faisceau) de l'Université de Wisconsin. En faisant l'analyse des movements de la mâchoire, les valeurs de magnitude et réglage syllabique étaient evalués pour inférer un enchaînement linéaire des pouls syllabiques pour représanter l'organisation rythmique de la énonciation. Les résultats préliminaires suggérent non seul que la magnitude de la syllabe corrigée augmente par la correction du chiffre, mais aussi qu'il y a en plusieurs cas quelque augmentation systématique de la magnitude de la syllabe, à la fois pour le chiffre corrigé et pour les autres chiffres de la même énonciation, quand la même correction est répétée. La variation considérable parmi les parleurs est observée et discutée sou forme de la magnitude syllabique et les modèles de réglage.

Keywords: C/D model; Articulatory gestures; X-ray microbeam; Jaw movement; Syllable magnitude; Syllable timing; Rhythm; Prosody; Emphasis; Repeated correction

1. Introduction

Emotion can affect speech in terms of changes in temporal organization, intensity, articulation or formants (stationary and dynamic characteristics), and voice quality, including F0 and voice source spectral envelope (Fant, 2000; Laver, 1994; Levelt, 1989; Williams and Stevens, 1972). Studies involving human and computer interaction by Oviatt et al. (1996) and Levow (1998) show that speech in error corrections tend to have increased utterance duration. Bell and Gustafson (1999) also dealing with human–computer interactions found that repetition of utterances leads to an increase in utterance duration, along with word hyperarticulation or placement of contrastive focus.

This paper addresses changes that occur in the rhythmic organization of an utterance due to the utterance condition, which may be labeled partly emotional, perhaps irritated, along with emphasis in conjunction with error corrections, in conversational speech during dialogue exchanges between an experimenter and a subject (Erickson et al., 1998). The theoretical framework underlying the data analysis is the C/D model (Fujimura, 1992, 1994, 2000), which assumes that the rhythmic organization of an utterance can be represented phonetically as a series of syllable and boundary pulses. These pulses vary in magnitude representing an abstract prosodic strength of each syllable. An abstract syllable duration is proportional to the magnitude of each syllable. The duration of each boundary also varies continuously and is assumed to be proportional to the magnitude of the boundary. The computed syllable-boundary pulse train representing the rhythmic pattern of the utterance defines a phonetic metrical grid.

The observed articulatory movement, in particular jaw opening, if the syllables contain a fixed low vowel as in our experimental material, seems to reveal the rhythmic organization, i.e., the stress pattern, of the utterance. Erickson et al. (1999) discuss preliminary data for different vowels, supporting the general concept of jaw opening representing the syllable magnitude. Such jaw opening control is accompanied by an enhancement of the inherent vocalic gestures relative to jaw position, when the syllable is emphasized. Considerable speaker idiosyncrasy with respect to prosodic strategy is noted in the current results.

1.1. The C/D model

The C/D model (Fujimura, 1992, 2000) deviates basically from the classical segment concatenation and coarticulation model (Lindblom, 1963; Farnetani, 1997). It uses syllables, rather than phonemic segments, as the minimal concatenative units, representing each syllable's internal structure directly in terms of phonological features and corresponding phonetic gestures. The model represents speech signals as an organization of articulatory events, by a derivational computational process of phonetic implementation. The input specification assumed for the process of mapping from an abstract representation of an utterance to concrete articulatory or acoustic signals includes, in addition to the phonological representation of the linguistic form, specifications of various utterance conditions as system parameters, as well as

paraphonological, discourse-related specifications for local prominence control.

Compared to other current models of phonetic organization, in particular, Articulatory Phonology (AP) (Browman and Goldstein, 1992) which adopts articulatory gestures at the lexical specification level, the C/D model (C/D) maintains the strict separation of phonology from phonetics. However, C/D assumes the phonetic system to be (parametrically) dependent on the particular language. C/D also assumes that phonetics handles symbolic as well as numeric variables inter-mixed in the computational process (Fujimura, 1992). C/D attempts a comprehensive representation of utterance characteristics in realistic conversational speech including emotional expression (Fujimura, 2000).

Tonal specifications for lexical and phrasal accent features are implemented in C/D as one of the melodic aspects of what is called the base function, separately from the representation of the skeletal pattern as the syllable-boundary pulse train. From this point of view, intonation patterns observed in the form of F0 contours, as discussed in a number of publications (for example, Pierrehumbert, 1980; Pierrehumbert and Beckman, 1988; Fujisaki, 1992), reflect both the stress pattern, as its default physiological manifestation in F0, and feature-specified tonal control. The tonal feature specification, as in Japanese lexical accent and some English phrasal accent, is realized as a melodic aspect of the base function. Other melodic aspects of the base function include vowel-to-vowel articulatory flow of movement, and switching between voiced and unvoiced phonetic status (Fujimura, 2000).

1.2. Rhythmic organization and emotion

Previous studies have shown that the syllable magnitude, as reflected in the amount of jaw opening, increases when additional prominence is attached to a word for discourse reasons such as contrastive emphasis (see Westbury and Fujimura, 1989; Erickson, 1998; Erickson, 2002; Erickson et al., 2000; Maekawa and Kagomiya, 2000—for the case of Japanese). Also, it has been reported (Spring et al., 1992; Erickson et al., 1998), based on the same data base as used in the current study, that, as the speaker becomes more irritated according to a listener's perception, the jaw tends to become more open. In these studies, emotion–perception tests were run for each sentential utterance taken out of its dialogue exchange context to see whether listeners perceived more irritation [1] for utterances produced as the speaker repeated the same correction several times. The results indicated that listeners actually tended to perceive irritation often in the same statement that was repeated many times in the same dialogue (see below). Mitchell et al. (2000) studying jaw movements indicated that not only does the magnitude of the corrected syllable change with repeated correction and perceived irritation, but also the magnitude and occurrence pattern of boundaries change, thus suggesting phonetic phrasal reorganization.

The current study does not relate observed articulatory characteristics to perceived emotion of the utterances per se. Instead, articulatory patterns are related to the number of repeated corrections the subject had to give in each dialogue, along with the intended correction of the particular digit in the three-digit sequence of the street address number. The position of the corrected digit in the digit sequence is also considered in analyzing the jaw opening pattern. A computational algorithm, developed in a previous study (Fujimura et al., 1998) was revised to process all data of the spontaneous dialogue Pine Street database (Erickson et al., 1998) to compute the hypothetical syllable magnitude and duration values for four speakers (two male and two female). This resulted in a quantitative evaluation of occurrences and magnitudes of boundaries between contiguous digits in different intraphrasal positions, i.e., between digits 1 and 2 and between 2 and 3 (Mitchell et al., 2000). In the current paper, however, boundary data are not directly described. Also, timing of each syllable pulse was derived by jaw opening time functions, simplifying the analysis algorithm used in the previous report (Mitchell

[1] Subjects more often chose "irritated" as the perceived quality among several labels.

et al., 2000). The current study focused on an initial exploration of speaker characteristics using all utterances that are currently available.

2. Experimental method

2.1. Data acquisition

Articulatory data with acoustic signals were recorded from 4 (2 male and 2 female) young native speakers of American English, who were students at the University of Wisconsin and who had no previous awareness of the research in progress. The experimenter (Donna Erickson) conducted a dialogue with each subject, whose articulation was recorded together with the acoustic speech signal. A dialogue used "five", "nine" and "pine" as the target words in a street address (e.g., 'five-nine-five Pine Street'; only the digits, not pine, were analyzed in this study). The dialogue was designed to allow the subject to be flexible in responding, always conveying the message of a given street address as instructed for each dialogue. Note that the same palatalized diphthong, /aJ/, was used for all the key words, in order to directly observe prosodic effects free from vowel identity effects. A previous paper (Fujimura et al., 1998) reported on a pilot study using sample utterances drawn from the same database. A preliminary result of the current study, pertaining only to Speaker 3, was reported in ISCA2000 Proceedings (Mitchell et al., 2000); the size of data for analysis has been expanded considerably producing more reliable data as a basis for further detailed studies, some of which will be reported elsewhere.

Acoustic and articulatory recordings were made at the Microbeam Facility of the University of Wisconsin, Madison (Nadler et al., 1987; Westbury, 1994). Spherical gold pellets (2.5–3 mm in diameter) were affixed to the tongue tip (several millimeters posterior to the extended tip) and two more flesh-points on the tongue surface in the middle and posterior portions, the lower lip (at the vermilion border) and mandible (the lower incisor). In addition, pellets were affixed to the bridge of the nose and to the anterior surface of the maxillary incisor as references. Relative to using these two reference pellets, the coordinate system was normalized in the preprocessing of the raw data by computational translation, rotation and amplification. This made it possible to correct for head movement during the utterance, producing midsagittal pellet position data fixed relative to the maxillary occlusal plane at a set distance from the X-ray pinhole lens (Westbury, 1994; see also Fujimura et al., 1973; Kiritani et al., 1975).

From each of the four speakers, 65–107 utterances were elicited in the experiment. The actual number of utterances obtained varied somewhat, because each dialogue was recorded within a fixed length of time (25 s) of the X-ray microbeam exposure, accommodating as many exchanges of correcting utterances as occurred. The subject's response varied from dialogue set to dialogue set, but the target phrase always contained the instructed digit sequence, which was one of the following three digit sequences: '595', '959' or '559'. 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. The subjects were instructed beforehand that the partner of the dialogue, seated invisibly, might need corrections because she was in a noisy communication environment. The subject would reply to the experimenter's repeated question, in which she deliberately mistook the same one of the three digits. A dialogue set consisted of several question–answer exchanges; the subject's first utterance was always the street address statement as displayed on the monitor screen (the reference utterance). A typical dialogue with one speaker is given below. In this dialogue set, the speaker is responding to the elicitor's "misunderstanding" of the initial digit, noted with capital letters in the subject's response.

Dialogue Set 6, Speaker 1 (female)
Exchange 0 (Reference). DE: Where do you live?
S1: I live at 959 Pine Street.
Exchange 1. DE: 559 Pine Street?
S1: No, NINE 59.
Exchange 2. DE: I'm...I'm sorry, I'm not hearing you... 559 Pine Street?
S1: It's at NINE 59.

Table 1
Inventory of the data, showing the number of dialogues, the total number of exchanges, and the number of digits per speaker

Speaker	# of dialogues	# of exchanges	Digits	# of digits
1 (female)	11	43	Five	69
			Nine	60
2 (male)	10	45	Five	76
			Nine	59
3 (male)	14	72	Five	123
			Nine	93
4 (female)	15	73	Five	125
			Nine	94

Exchange 3. DE: I'm sorry—that was 559 Pine Street, right?
S1: No, not...it's NINE 59.
Exchange 4. DE: OK, 559 Pine Street?
S1: Not, 559, NINE 59.

Similar sets of dialogue occurred a few times (exact number being variable due to occasional pellet mistracking of the microbeam system), using the same digit sequence as the correct statement and the same erroneous digit sequence repeated by the experimenter. The digit to be corrected, among the three digits in sequence, was controlled for each correct (target) digit sequence, to appear in initial, middle, or final digit position. Thus the condition of the three-digit sequence corrected on any one of the three digits, typically heard with an enhanced prominence, was used in the data of each speaker 0–3 times, [2] resulting in a total number of dialogues with the same digit sequence to be 2–9. The total number of dialogues using any of the three different digit sequences was 11–15 for each speaker (see Table 1). The number of exchanges within each dialogue set varied from 4 to 7 including the reference exchange. The total numbers of sample digits recorded are tabulated in Table 1. Some other dialogues that were not used for the current analysis were also recorded in between the relevant dialogues as experimental foil.

[2] Speakers 1 and 3 did not have the dialogues with the sequence 959 with the last digit corrected.

2.2. Inference of syllable magnitude and timing

All articulatory data from the spontaneous dialogue Pine Street database for the present study were analyzed using a version of the MATLAB-based program UBEDIT (by Bryan Pardo). Statistical analyses were performed using SPSS in combination with Microsoft Excel 97.

UBEDIT displays the acoustic signal and time functions for horizontal and vertical movements of data pellets placed on the subject's articulators. In this study, mandible height was the primary object for measurement, while other pellet positions for lower lip, and tongue tip, along with the acoustic waveform, were always displayed and the acoustic signal was played as needed during data analysis. The program measures the position value and timing of the minimum vertical position of the mandible for each target syllable (for 'five' or 'nine' only in the main three-digit sequence in each dialogue), along with the maximum vertical positions for the crucial syllable onset/offset articulators (lower lip for 'five' and tongue tip for 'nine') for facilitating semi-automatic data processing as described below. This identification and labeling was performed by careful visual and auditory inspection of each utterance. The time value of each jaw minimum was automatically determined, under interactive visual inspection by smoothing the recorded time function using a 10 data point-wide gaussian window centered on the current sample. The zero-crossing points of the first time derivative were then taken from the smoothed signal for identifying the position extremum.

The magnitude (height) of each syllable pulse is assumed to be proportional to the maximum jaw opening value, measured as the distance from the occlusal plane to the minimum mandible position (Fujimura et al., 1998). Using these jaw position measurements, UBEDIT automatically constructs a syllable pulse for each syllable at the time of the jaw minimum. [3]

[3] An option of the program computes a revised time value for the syllable pulse based on a revised iceberg algorithm (see Mitchell et al., 2000 for more detailed analyses of syllable and boundary duration).

The program also has an option to construct a syllable triangle, centering around each syllable pulse in order to compute articulatory syllable duration (Fujimura et al., 1998; Mitchell et al., 2000). Based on this syllable triangle, which defines an abstract syllable duration, gaps between consecutive syllable triangles are interpreted as boundary magnitudes (Fujimura et al., 1998; Mitchell et al., 2000). The present paper concentrates on the discussion of syllable pulse magnitude and approximate timing information. This magnitude–time information of the pulse train represents the rhythmic pattern of the pertinent portion of each utterance under varied prosodic conditions, reflecting the intended correction of one of the digits and also repetition of the same correction as requested by the dialogue partner.

3. Results

3.1. Changes in syllable magnitude for corrected digits

To remove the direct effects of digit position in the phrase and of the digit identity ('five' or 'nine'), the jaw opening value of the reference utterance was subtracted from the corresponding jaw opening value in each correcting utterance of the same dialogue. In other words, the jaw opening value of digit i ($i = 1, 2, 3$) in the reference utterance was subtracted from the jaw opening value of the corresponding digit i in each repeated correction. The digits in the reference utterances always had the magnitude value of "0". The digits in the corrected utterances would have a positive or negative value. These incremental values were then used to analyze the effect of digit correction and repeated correction on syllable magnitude.

Speakers were studied separately to examine inter-speaker variability. Analysis of jaw opening values only for reference utterances showed that on average all speakers had a relatively large jaw deviation for the initial digit in comparison with middle and final digits. According to the assumption of this study, jaw opening is proportional to syllable magnitude; therefore, it can be said that all speakers tend to use a large syllable magnitude at the beginning of a phrase. This is in conformity with Keating's (1995) suggestion that the initial part of any syntagmatic unit is generally strong. However, speakers varied with regard to the second and third digits of the sequence. Speakers 1 and 4 tended to use the least amount of jaw opening for the middle digit, while Speakers 3 and 2 for the final digit (Table 2). To end a phrase, some speakers may gradually decrease syllable magnitude, while other speakers may increase it.

Table 2
Mean syllable magnitude (jaw opening) for all digit positions in reference utterances (in mm)

Speaker	Digit position	N	Mean (±SE)
1	Initial	12	14.4 (0.5)
	Medial	12	11.9 (0.3)
	Final	12	13.2 (0.3)
2	Initial	10	14.7 (0.3)
	Medial	10	14.5 (0.6)
	Final	10	13.0 (0.4)
3	Initial	14	13.8 (0.3)
	Medial	14	11.4 (0.3)
	Final	14	10.9 (0.3)
4	Initial	15	11.7 (0.4)
	Medial	15	09.5 (0.3)
	Final	15	10.5 (0.3)

Mean incremental values were analyzed also to study the effect of digit correction on first correction utterances and repeated correction utterances. The syllable magnitude for all digits in the repeated correction was significantly larger than the digits in the reference utterances (ANOVA, $F_{(2,708)} = 21.18$, $p = 0.001$). Further, the corrected (emphasized) digits on the average had a larger syllable magnitude than the uncorrected digits (unemphasized) consistently for all speakers (see Fig. 1).

Fig. 1, using histograms, shows how jaw opening increases for individual speakers as an effect of contrastive (i.e., correcting) emphasis. Bars represent the mean incremental value of jaw opening for all emphasized digits and unemphasized digits across all repetitions. These data suggest that all speakers tend to use large syllable magnitudes to effect corrections in their utterances. This is true, in most cases for the entire phrase of correcting ut-

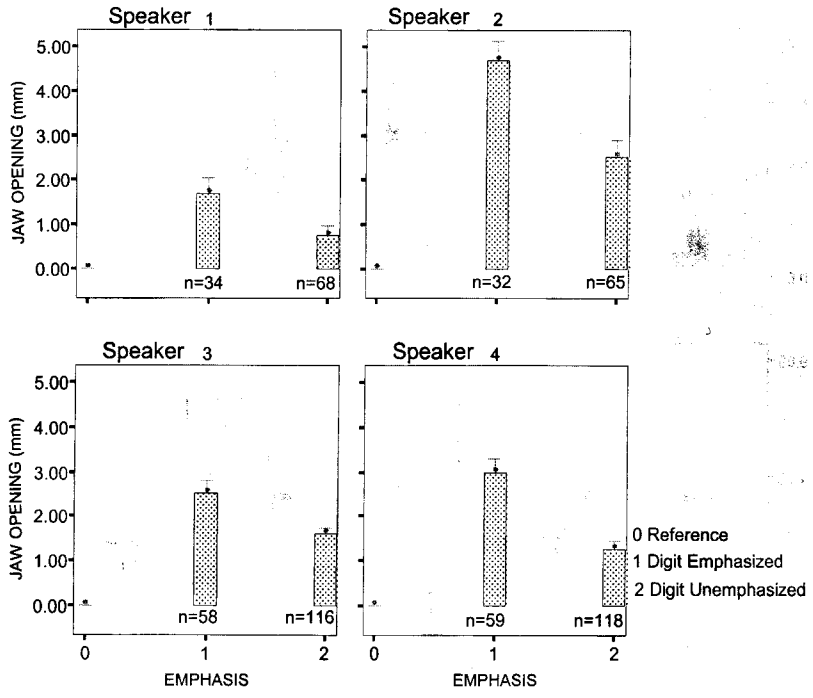

Fig. 1. Mean incremental values of jaw opening (i.e. syllable magnitude) for emphasized and unemphasized digits across all repeated corrections separated by speakers. Error bars represent mean ± 1.0SE.

terances including other digits that are not corrected. The corrected digit, in particular, is made the most prominent.

3.2. Changes in syllable magnitude due to repetition of a correction

Incremental values were also examined for evaluating the pattern of syllable magnitudes due to repetition of the same correction in a dialogue. The results are displayed in Fig. 2, where bars represent mean incremental values as the number of repeated corrections increases for the emphasized and unemphasized digits. The error bars are the standard error of the mean. Data are pooled for all dialogues within a speaker. The graphs reveal that all speakers, with the exception of Speaker 1, clearly show an increase in jaw opening, on the average across all dialogues, as the number of repetition increases. A comparison of means, using ANOVA, revealed that this increase in jaw opening as a result of repetition of corrections was significant when all the speakers were pooled together ($F_{(5,705)} = 3.68$, $p = 0.003$). This increase in jaw opening was manifested in both the emphasized digits and the unemphasized digits (Fig. 2) in the correcting utterances. Speaker 1 generally decreased jaw opening with repeated corrections, except for the very last exchange. No obvious explanation could be found for this variation in Speaker 1.

3.3. Articulatory metrical grid

Fig. 3 illustrates an example of two syllable pulse trains of the 3-digit sequence '959' in the first and second exchanges of Speaker 1, Dialogue 6, which is the second dialogue recorded in the experiment for this speaker. In this figure, the thick bars are for the reference utterance and the thin bars are for the first correction, i.e., the second exchange, with the initial digit corrected. The ordinate represents the jaw opening in millimeters and the abscissa represents time relative to the

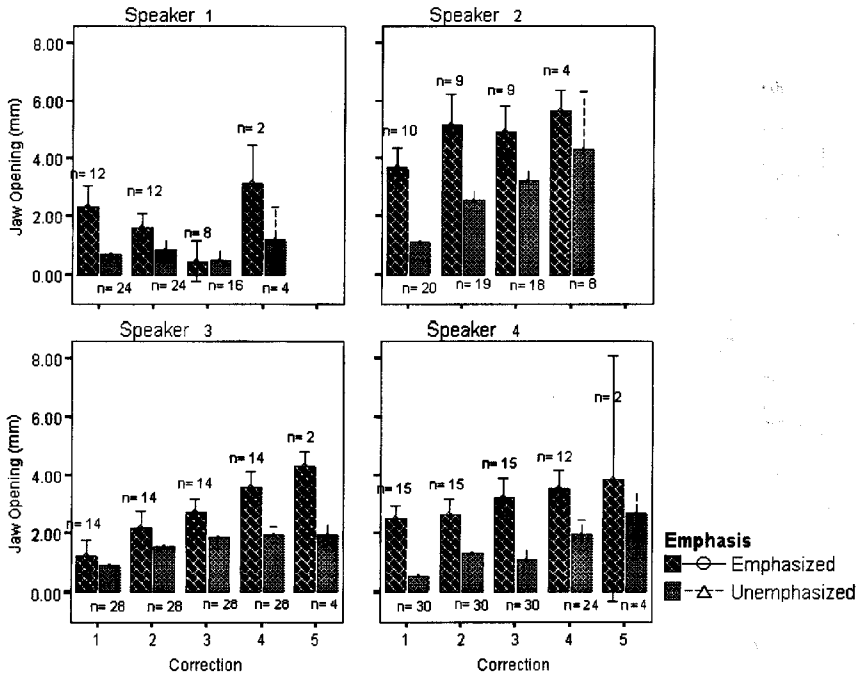

Fig. 2. Mean incremental values across all digits for all repetitions of correction separated by speakers and emphasis condition. Error bars represent mean ± 1.0SE emphasized condition refers to the digits that were corrected, and unemphasized refers to the digits that were not corrected in the repeated corrections.

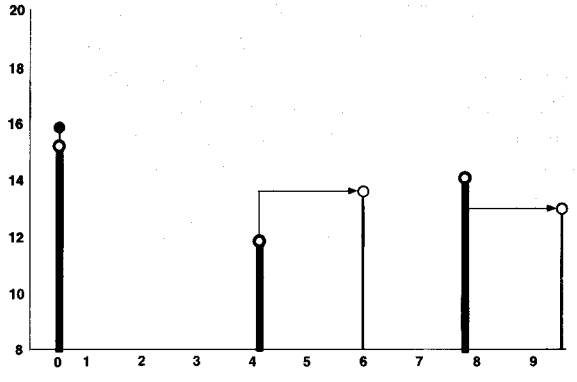

Fig. 3. Two metrical grids derived from jaw movement, for exchanges 1 and 2 of Speaker 1, Dialogue 6. The thick lines represent the reference utterance and the thin lines the first correction. On the y-axis is syllable magnitude and on the x-axis is relative time the pulses occur. The arrows show the change effected for the middle and final digit as a result of first correction.

initial syllable pulse in deciseconds (100 ms). Thin horizontal arrows and a vertical line segment are drawn to assist comparing the corresponding syllable pulses in the two utterances. The tip of each pulse is marked by a circle to show the magnitude-time value of each syllable pulse. The corrected digit is given a shaded circle. These syllable pulse trains (thick and thin) may be interpreted as phonetic metrical grids with timing of syllables indicated by position along the abscissa; they show how the prosody (rhythm) of the utterance changed due to correction (in this case on the initial digit). The initial and second digits are both strengthened by the correction of the first digit. This effect is observed, according to the C/D model (Fujimura, 2000; Fujimura et al., 1998), as increases in both the magnitude values of the initial and middle syllable pulses and, the time interval between the two pulses.

In this particular utterance example, the final syllable pulse is decreased in magnitude as the result of the initial digit correction. Interestingly, the time interval between the second and third pulse is not altered much. As can be seen in Fig. 4, comparing the second panel (first correction) with the

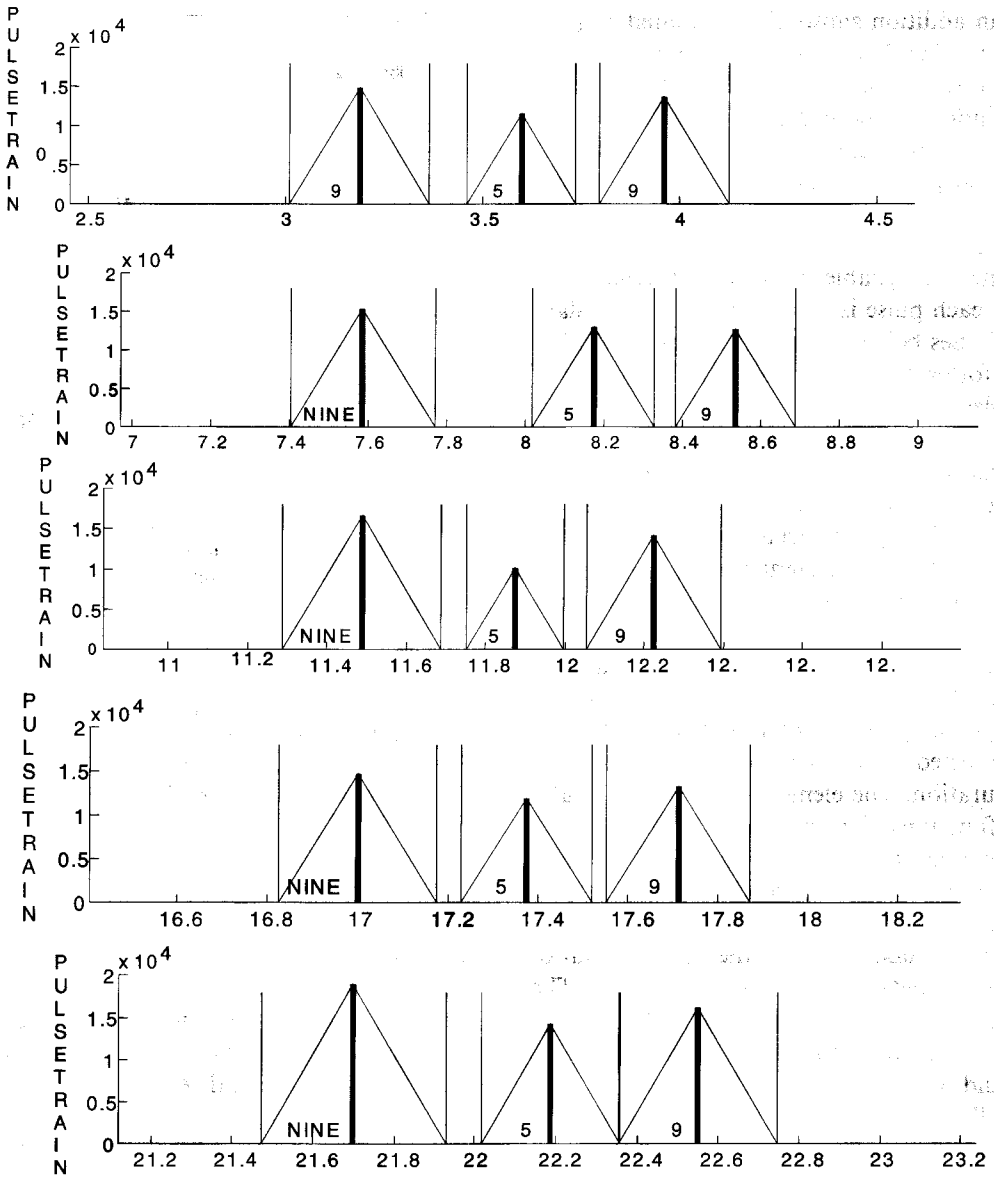

Fig. 4. Pulse-trains for Speaker 1, Dialogue 6, as seen on UBEDIT. Each frame represents each exchange of the speaker, where the reference is the top frame and the last correction is the bottom frame. One can see that in the last frame (fourth repeated correction) there are contiguous syllable triangles corresponding to a no boundary situation.

top panel (reference) for the same utterances as in Fig. 3, the UBEDIT interprets that the second syllable triangle is enlarged while the third syllable triangle is reduced by the first digit correction. The two effects cancel each other in the evaluation of the time interval between the second and third syllable pulses. In contrast, the time interval between the first and second syllable pulses (of the second panel) is considerably enlarged by the initial digit correction. This must be ascribed to an enlarged boundary between the two digits as well as the increase of both the first and second triangles.

Fig. 4 in addition shows the computed syllable triangles for all exchanges 1 through 5, from top to bottom, in Dialogue 6 of Speaker 1. The label "NINE" indicates the corrected first digit of the phrase '959'. This figure explains how the timing of each syllable pulse, representing the time of occurrence of mandible height minimum in this study, is interpreted by UBEDIT. Each thick vertical line is a syllable pulse. Constructed from the top of each pulse is a "syllable triangle"; slant "shadow" lines being drawn downward to the left and right for each syllable, such that the angles of all triangles are identical throughout the dialogue, [4] and the angle has the maximum value without causing any overlapping of consecutive triangles (Fujimura et al., 1998; Mitchell et al., 2000). (Note: The syllable triangle is assumed to be symmetric left to right (Fujimura et al., 1998). This does not imply that onset and coda consonants have articulatory behaviors in a mirror image, which is known not to be the case (Sproat and Fujimura, 1993; Krakow, 1999). It simply means that the syllable pulse, to represent the whole syllable, is erected at the center of the computed syllable duration. The elemental gestures (impulse response functions) for onset and coda are generally not symmetric in the C/D model.)

The resultant gaps between consecutive triangles within a digit sequence are interpreted to represent the boundaries, and the time interval of each gap represents the boundary magnitude. The last exchange (bottom frame, Fig. 4) shows a condition where no boundary occurred between digits 2 and 3. In other words, this pair of contiguous syllables constitute the critical syllable sequence that determined the shadow angle (the angle formed by the vertical pulse and the slant lines to left and right, the triangle being assumed to be symmetrical).

Fig. 5 summarizes the syllable pulse magnitude–time information for the same dialogue shown in

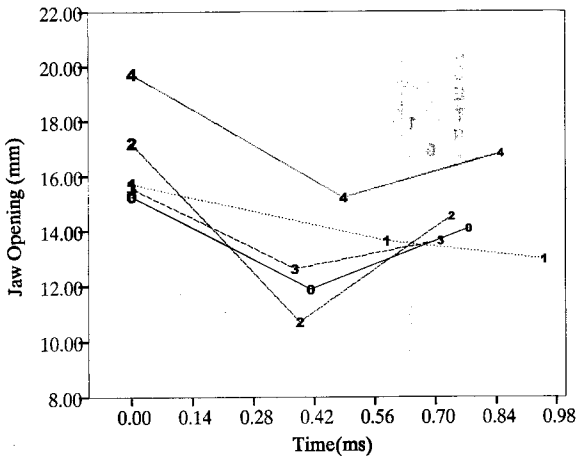

Fig. 5. Syllable pulse magnitude (y) versus time (x) for exchanges 1–5 for Speaker 1, Dialogue 6. The first grouping of integers represents the mean syllable magnitudes for the first digit ('9' or 'NINE'); the second grouping of integers represents the mean syllable magnitudes for the second digit, '5'; and third grouping, the mean syllable magnitudes for the third digit, '9'. The different integers indicate the exchange number, where '0' is the reference utterance and '4' is the fourth correction (fifth exchange in the dialogue). The largest integers indicate the digits in the initial position. The smallest integers indicate the final digit position and the intermediate size integer the middle digit in the three-digit sequence. In this dialogue the initial digit is the corrected digit.

Fig. 4. Subsequent figures use the same format of representing syllable magnitude–time relations as scatter plots of our database. Note different integers are used to indicate the exchange number, integer 0 for the first exchange (reference) and number 4 for the fifth exchange (i.e., the fourth correction). Change in integer size is used to indicate digit position, such that the initial digit has the largest integers and the final digit has the smallest integer size. For this dialogue the initial digit was emphasized. Generally, with some exceptions, the more the correction is repeated, the more the syllable magnitudes and time values tend to increase. Note also, however, that in this dialogue the second digit is actually reduced when the first digit is corrected for the second time (the integer 2 in the medium size). The corrected digit itself (integer 2 in the largest size) is considerably higher in magnitude than in the reference condition (see also Fig. 4). A previous publication dis-

[4] The current version of UBEDIT allows the user to set the shadow angle so that it is constant throughout different dialogues. This feature turns out to be necessary given that some of our speakers tended to insert boundaries always between consecutive digits in the sequence throughout some dialogues.

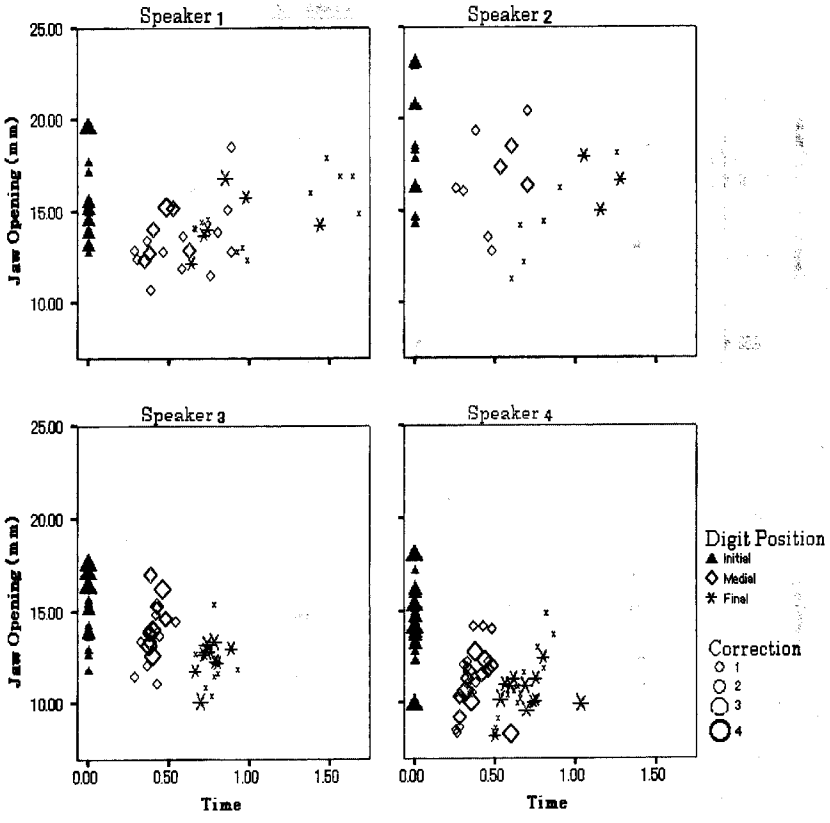

Fig. 6. Jaw opening values (y) plotted against time relative to the timing of the first digit for all initial digit corrections. Filled marks represent corrected digits. The two types of syllables (five and nine) are pooled. Correction number is indicated by size of the symbol, such that, the reference has the smallest symbol and the fourth correction, the largest symbol.

cussed effects of contrastive emphasis increasing the drop in prominence from the directly affected syllable to the immediately following syllable in some speakers (Erickson, 1998).

Figs. 6–8 compare syllable magnitude-time patterns for different speakers in all dialogues. Digit position is indicated by different shapes of symbols. Varying the size of the symbols used in the graph differentiates repeated corrections with the later correction being larger in size. The digit that is emphasized in Figs. 6–8 has been given the filled symbol. Fig. 6 shows the effect of the initial digit being corrected separately for all the speakers as the number of correction repetition increases. The initial digit (corrected digit, represented by the filled symbol) has the largest jaw opening for all speakers. All speakers (Fig. 6) also tend to show a larger time interval between the initial and middle digits in comparison to the time interval between the middle and final digits, regardless of syllable magnitudes. This implies that speakers tend to insert or increase the strength of the boundary between the initial and middle digits, when the same correction is repeated, if the correction is made on the first digit. It may be taken to suggest that speakers tend to place a boundary after the corrected digit in general. However, when other digits are corrected, as we see below, there is some inter-speaker variability for the treatment of boundaries. While most speakers also show clear temporal separation of the middle and final digits when the first digit is emphasized, Speaker 1 does not follow this trend. This is another example of speaker variability.

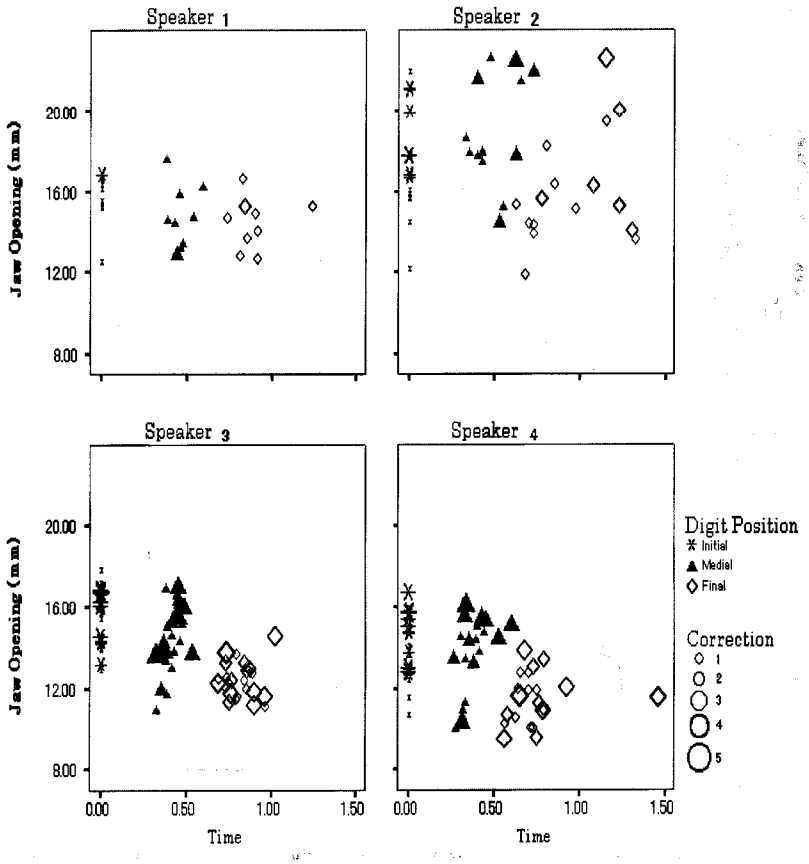

Fig. 7. Jaw opening values (y) plotted against time relative to the timing of the first digit for middle digit corrections. Filled marks represent corrected digits. The two types of syllables (five and nine) are pooled. Correction number is indicated by size of the symbol, such that, the reference has the smallest symbol and the fifth correction, the largest symbol.

Fig. 7, for the middle digit corrected, shows that Speakers 3 and 4 clearly place a large boundary between the middle and final digits after the corrected digit, while Speakers 1 and 2 tend to have a larger boundary between the middle and initial digits before the corrected digit. With regards to the middle digit correction, speakers thus seem to opt for different phrasing strategies. These graphs also confirm that speakers have greater jaw opening for the corrected digit and it tends to increase as the number of correction increases.

Fig. 8 indicates the effects of final digit correction. Speakers show a large jaw opening for the final digit as well as a large time interval between the middle and final digits. This suggests that speakers tend to place a boundary before the emphasized digit, possibly as well as after it, if the digit is at the final position of a three-digit sequence. Although there is much variation in natural speech there also seems to be some constant features.

4. Discussion

In this article, we primarily focused on magnitude and time values, both determined by jaw movement patterns of a relatively large amount of articulatory data, according to the theoretical framework of the C/D model (Fujimura and Williams, 1999; Fujimura, 2000; Mitchell et al., 2000). Some detailed evaluation of syllable timing requires a more complex data processing algorithm, some of which was discussed in previous presentations but was not feasible to include in this

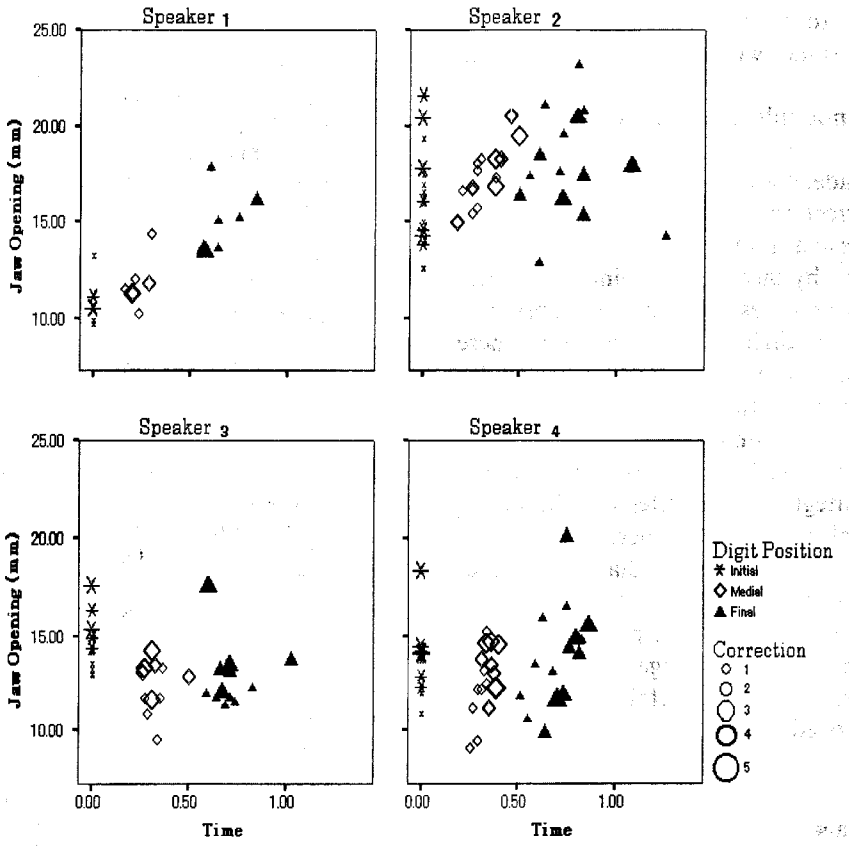

Fig. 8. Jaw opening values (y) plotted against time relative to the timing of the first digit for all final digit corrections. Filled marks represent corrected digits. The two types of syllables (five and nine) are pooled. Correction number is indicated by size of the symbol, such that, the reference has the smallest symbol and the fifth correction, the largest symbol.

processing of more extensive data covering different speakers at this time. For this reason, we did not discuss quantitative characteristics of boundary durations, and consequently phonetic phrasing patterns. Our spontaneous Pine Street dialogue data seem to provide a rich source of information for such discussion. Two phonetic factors: (1) strong coarticulatory smoothing of gestures in time series and (2) phrase-final elongation effects, seem to characterize some of the utterances observed, particularly those produced by Speaker 2. The latter, in particular, is an unexplored domain of the C/D model from a quantitative point of view. It is a topic deserving future investigations. Speaker 2, a male speaker, seemed to reveal particularly strong prosodic variation, which challenged our automatic iceberg algorithms (Fujimura et al., 1998; Mitchell et al., 2000) for identifying all of the syllables and evaluating their magnitude and timing accurately. Since it was not possible to apply the currently implemented iceberg algorithm for all utterances across all speakers, we decided to evaluate syllable pulse magnitude and time values directly from jaw minima for this study.

Although these were recorded in a laboratory setup and, as such, not completely natural as free dialogue, the particular acquisition technique, which was introduced by Erickson in the microbeam laboratory environment, extending a previous experimental design for "read" dialogues (Westbury and Fujimura, 1989), seems to have worked remarkably well. It has led us to findings about changes in syllable magnitude

and timing due to widely varied prosodic conditions of the same words by using repeated corrections.

In summary, not only do most speakers indicate correction in their communication by increasing syllable magnitude, they also continue this trend for repeated corrections in an enhanced form. Our data also suggested that speakers indicated repeated correction by inserting or reinforcing phonetic phrase boundaries between the corrected digits and adjacent digits in the sequence. These boundaries need not be accompanied by an increased syllable magnitude. To the extent that these data try to simulate natural speech as much as possible, we can conclude that speakers may use a variety of strategies for manifesting the act of correction and also, sometimes, perhaps responding to an unusual insistence of the dialogue partner to keep repeating the same correction. The C/D model seems to work effectively, providing a new descriptive framework for a large variety of prosodic conditions, while many details remain to be added and improved.

Acknowledgements

Part of this work was supported by research grants from the National Science Foundation, USA (BCS-9977018/SBR-9511998, PI: O. Fujimura) and a gift from ATR, Japan (PI. O. Fujimura). Corey Mitchell played a principal role in an earlier phase of this project. J.C. Williams contributed valuable discussion and suggestions; some based on her own experience with the spontaneous Pine Street data.

References

Bell, L., Gustafson, J., 1999. Repetition and its phonetic realizations: investigating a Swedish database of spontaneous computer directed speech. In: Proc. ICPHS99, San Fransisco, pp. 1221–1224.

Browman, C.P., Goldstein, L.M., 1992. Articulatory phonology: An overview. Phonetica 49, 155–180.

Erickson, D., 1998. Effects of contrastive emphasis on jaw opening. Phonetica 55, 147–169.

Erickson, D., 2002. Articulation of extreme formant patterns of emphasized vowels. Phonetica 59, 134–149.

Erickson, D., Fujimura, O., Pardo, B., 1998. Articulatory correlates of prosodic control: emotion and emphasis. Lang. Speech 41, 395–413.

Erickson, D., Fujimura, O., Dang, J., 1999. Articulatory and acoustic characteristics of emphasized and unemphasized vowels. J. Acoust. Soc. Amer. 106, 2241.

Erickson, D., Maekawa, K., Hashi, M., Dang, J., 2000. Some articulatory and acoustic changes associated with emphasis in spoken English. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP-2000), Beijing, Vol. 3, pp. 247–249.

Fant, G., 2000. The source-filter frame of prosody. Phonetica 57, 113–127.

Farnetani, E., 1997. Coarticulation and connected speech processes. In: Hardcastle, W., Laver, J. (Eds.), The Handbook of Phonetics. Blackwell Publishers, Cambridge, MA, pp. 371–404.

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

Fujimura, O., 1994. C/D model: A computational model of phonetic implementation. DIMACS Series in Discrete Mathematics and Theoretical Computer Science (Amer. Math. Soc.) 17, 1–20.

Fujimura, O., 2000. The C/D model and prosodic control of articulatory behavior. Phonetica 57, 128–138.

Fujimura, O., Williams, J.C., 1999. Syllable concatenators in English, Japanese and Spanish. In: Fujimura, O., Joseph, B., Palek, B. (Eds.), Item Order: Proc. LP'98. Cambridge University Press, Cambridge, pp. 40–51.

Fujimura, O., Kiritani, S., Ishida, H., 1973. Computer controlled radiography for observation of movements of articulatory and other human organs. Comp. Biol. Med. 3, 371–384.

Fujimura, O., Pardo, B., Erickson, D., 1998. Effect of emphasis and irritation on jaw opening. In: Proc. European Speech Community Association Conf. on Sound Patterns of Spontaneous Speech: Production and Perception (ESCA), Aix-en-Provence, France, pp. 23–29.

Fujisaki, H., 1992. Modeling the process of fundamental frequency contour generation. In: Tohkura, Y., Vatikiotis-Bateson, E., Sagisaka, Y. (Eds.), Speech Perception, Production and Linguistic Structure. IOS Press, Amsterdam, pp. 313–326.

Keating, P., 1995. Segmental phonology and non-segmental phonetics. In: Proc. Internat. Congress of Phonetic Sciences (ICPhS-1995), Stockholm, Vol. 3, pp. 26–32.

Kiritani, S., Itoh, K., Fujimura, O., 1975. Tongue–pellet tracking by a computer-controlled X-ray microbeam system. J. Acoust. Soc. Amer. 57, 1516–1520.

Krakow, R.A., 1999. Physiological organization of syllables: a review. J. Phonetics 27, 23–54.

Laver, J., 1994. Principles of Phonetics. Cambridge University Press, Cambridge.

Levelt, W.J.M., 1989. Speaking: From Intention to Articulation. MIT Press, Cambridge, MA.

Levow, G., 1998. Characterizing and recognizing spoken corrections in human–computer dialogue. In: Proc. COLING/ACL'98, pp. 736–742.

Lindblom, B., 1963. Spectrographic study of vowel reduction. J. Acoust. Soc. Amer. 35, 1773–1781.

Maekawa, K., Kagomiya, T., 2000. Influence of paralinguistic information on segmental articulation. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP-2000), Beijing, Vol. 2, pp. 349–352.

Mitchell, C.J., Menezes, C., Williams, J.C., Pardo, B., Erickson, D., Fujimura, O., 2000. Changes in syllable boundary strengths due to irritation. In: Proc. ISCA 2000, Newcastle, pp. 98–101.

Nadler, R.D., Abbs, J.H., Fujimura, O., 1987. Speech movements research using the new X-ray microbeam system. In: Proc. XIth Internat. Congress of Phonetic Sciences, Tallinn, Vol. 6, pp. 10–27.

Oviatt, S., Levow, G., MacEachern, M., Kuhn, K., 1996. Modeling hyperarticulate speech during human–computer error resolution. In: Proc. ICSLP'96, pp. 801–804.

Pierrehumbert, J., 1980. The phonology and phonetics of English intonation. Doctorial Dissertation, MIT, Distributed by: Indiana University Linguistics Club, Bloomington, Indiana.

Pierrehumbert, J.B., Beckman, M.E., 1988. Japanese tone structure. Linguist. Inquiry Monogr., 17.

Spring, C., Erickson, D., Call, T., 1992. Emotional modalities and intonation in spoken language. In: Ohala, J. (Ed.), Proc. Internat. Conf. on Spoken Language Processing, Alberta, Canada, pp. 679–682.

Sproat, R., Fujimura, O., 1993. Allophonic variation in english/l/and its implications for phonetic implementation. J. Phonetics 21, 291–311.

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.R., Fujimura, O., 1989. An articulatory characterization of contrastive emphasis. J. Acoust. Soc. Amer. 85 (S1), s98 (A).

Williams, C.E., Stevens, K.N., 1972. Emotions and speech: some acoustic correlates. J. Acoust. Soc. Amer. 52, 1238–1250.