ESCA 98 Aix en Provence
September 1998

EFFECT OF EMPHASIS AND IRRITATION ON JAW OPENING

Osamu Fujimura, Bryan Pardo, and Donna Erickson
Department of Speech & Hearing Science, The Ohio State University
Columbus, OH 43210-1002, U. S. A.
E-mail: fujimura.1@osu.edu, Fax: +1 614 292 7504

RESUME

Les traces des mouvements mandibulaires, observées au système des rayons X micro-faisceaux (x-ray microbeam), pendent que les sujets prenaient part aux dialogues, sont interpretées selon le modèle Convertisseur/Distributeur (C/D). L'experimenteur faisait sortir du sujet une adresse (559 ou 595 Pine St., selon l'essai), et puis faisait semblant de mal comprendre l'un ou l'autre des numéros à plusiers fois. Le sujet le corrigait chaque fois, avec le résultat des dialogues variés mais pareils. En utilisant le méthode "iceberg", le centre de chaque syllable est évalué, le faîte de pouls est déterminé selon l'amplitude excursive de la mandibule, et une "durée syllabique" est calculée au moyen de la construction d'un triangle syllabique, toutes ces opérations exécutaient pour un programme informatique. Les intervalles qui restent entre les domaines syllabiques contigus sont interprétés comme la portée de la borne syllabique. Une varieté de moyen de communiquer les corrections avec émotion est démontrée par des examples.

1. INTRODUCTION

According to the C/D model [1, 2, 3], the prosodic organization of an utterance can be represented by the magnitude distribution of a pulse train comprising a linear time series of syllable and boundary pulses. The timing of each syllable pulse is computed from the magnitude of the pulse relative to the magnitude and time of the immediately preceding syllable or boundary pulse. The magnitude-time pattern of such a syllable-boundary pulse train represents the skeleton of what we call the base function for a given utterance. Other aspects of the base function represent the phonetic status of syllable components (*i.e.* nucleus, onset, coda, each s-fix or p-fix) or of each boundary (*e.g.* phrase accent) in different physiological dimensions: vocalic control, voicing control, boundary tones, *etc.* Altogether, the base function comprehensively represents what we might consider as the generalized prosodic pattern of the utterance [3], which is implemented physiologically and physically in different muscular control devices including peripheral feedback loops in the speech production process. These processes utilize subcortical brain functions coordinated by cortical motor planning for phonetically implementing phonological specifications. materials.

Upon this base function, consonantal elemental gestures are superimposed. Their temporal characteristics including timing (relative to the controlling syllable pulse) are all fixed and stored as impulse response functions (IRFs); their amplitudes directly reflect the magnitude of the pertinent syllable pulse. In other words, each syllable pulse excites the set of IRFs for all the elemental gestures that are evoked by the distributor component of the C/D model for the pertinent syllable, according to the phonological feature specifications as evaluated by the converter [1, 2, 3].

The phonological features are assumed to be specified with a syllable-based underspecification scheme [2]. The magnitude patterns of the syllable and boundary pulses are computed by numerically augmenting the phonological metrical structure [4].

This paper discusses some preliminary experimental findings of jaw movement patterns as observed in a fairly natural dialogue situation. The data of this conversational speech material was designed and acquired by Donna Erickson using the X-ray Microbeam Facility of the University of Wisconsin, Madison [5]. This paper describes a new analysis method to interpret articulatory movement patterns and derive the syllable/boundary pulse train. Some preliminary results are shown about the role of boundaries that seem to give new insight into the nature of contrastive emphasis, in relation to conversational error correction, and also some expression of irritation.

2. SYLLABLE MAGNITUDE

According to the C/D model, the magnitude of each syllable pulse, as an abstract prosodic strength, should be reflected in a number of physical properties of the phonetic signals, physiologic, mechanical, or acoustic. Also, the theory predicts that there is a strong correlation among a number of measurable quantities of tautosyllabic events such as the excursion of the mandibular movement, extent of crucial flesh point movements for both consonantal and vocalic gestures in different parts of the syllable, as well as manifestation of tonal specifications.

Most apparently, perceptually and spectrographically, the temporal stretch (duration) of physical events, manifesting consonantal and vocalic feature specifications, must vary with the syllable magnitude [4]. As seen in the top panel of Fig. I, syllable pulse shadows on both sides of the pulse expands proportionally (*i.e.* keeping the angles constant) as the syllable pulse height increases. The extension of the shadow edges results in placing the elemental consonantal (including glide) gestures of the syllable margins further away from the center of the syllable [5]. The IRFs are excited by the *pocs* pulses [6] that are the same in amplitude as the syllable pulse but displaced in time to the outer corners of the syllable triangle.

The extent of mandibular excursion from the syllable initial margin to the nucleus and back to the final margin, appears to reflect the syllable magnitude most directly, if we set aside the so-called segmental effects on jaw opening, in particular its dependence on vowel height. In the current experimental design, we used the same vowel /aJ/ for the key words 'five', 'nine' and 'pine'.

Assuming this relation between the syllable duration (as defined by the temporal stretch of the pulse shadows) and the jaw excursion reflecting the abstract syllable magnitude, we interpreted typical examples of dialogue utterances in our conversational Pine Street data [6]. This figure pertains to an utterance by a male subject saying, in an exchange with the experimenter: 'FIVE five nine Pine Street', in reply to the experimenter's third repeated question deliberately mistaking the first digit asking if the address was '959 Pine Street'.

In the top panel, the first pulse around the time value 18.6 (second) is not a syllable spoken but reflects a silent jaw opening gesture preceding the utterance. The program identified the first local dip in the mandible height time function (MAN_{ly}), shown in the third panel. It evaluated the extent of excursion (deviation from the gradual down slope to the bottom of the local dip) and treated it as the pertinent syllable pulse height.

The next triangle represents a syllable /faJv/; the Lly time function (second panel) representing the vertical movement of the lower lip pellet shows clearly peaks with plateaus on both sides of the syllable nucleus dip. This movement of the crucial articulator for the margins of this syllable is labeled "FIVE" in this panel.

When this syllable identification is given, the program defines the temporal center of this syllabic movement pattern by evaluating the times for the downward and upward movements to cross a fixed (iceberg [7]) threshold height (shown by a horizontal line) and taking the mid point.

Similarly, the next syllable pulse for the second digit 'five' is created. The third digit in the phrase is 'nine', and the crucial articulator for the syllable margin is the tongue tip. The shadow angles are always kept constant. The bottom panel shows the time function of its vertical movement (TT_y), and a similar evaluation procedure of the syllable center provides the time position of the fourth pulse with a length determined by the mandibular excursion.

This algorithm for finding a representative time value for each syllable pulse was empirically found more reliable than simply finding the jaw minimum time for each syllable, even though it is effective only for demisyllables containing a place-specified obstruent consonant. As seen in this figure, jaw minima tend to be asymmetric partly due to the gradual drifting of the vertical position related to the phrasing of the utterance and partly due to phrase-final effects, possibly due to irritation.

This utterance in Fig. I sounds fluent without using much temporal cueing for contrastive emphasis. This is typical for this speaker, while other speakers in the experiment used different means of expressing emphasis as seen below (our data base contains many similar exchanges by each of four speakers, two male and two female). The emphasis placed on the first digit 'five' is perceptually obvious in this case as phonetic prominence. Correspondingly, the mandibular movement pattern (center panel) shows for the first digit a distinctly larger excursion than for other digits. This is reflected as the salience in size of the first syllable triangle (i.e. second from left in the top panel), in contrast with the last digit showing the highest jaw position minimum among the three digits (note that the 'Pine' following the three digits shows an even higher jaw minimum and also a very small syllable triangle). In this utterance, which shows some sign of resigned attitude toward the end of the dialogue, the mandibular position at the syllabic minima rose gradually and considerably toward the end of the utterance.

In computing the syllable durations, depicted in Fig. I, top panel, as the length of the bottom of each syllable triangle, the shadow angle used was symmetric and fixed for all syllables in this dialogue. The angle was automatically determined to optimize the contiguity of shadows allowing no overlapping. In this example, relatively small gaps between successive syllables are noted between the syllable domains thus defined.

3. BOUNDARY MAGNITUDE

Using this algorithm for determining syllable duration in our sense, we can evaluate quantitatively boundary strengths as the remainder of the time span of the utterance not accounted for by the computed syllable domains. Fig. II shows an example similar to Fig. 1 but pertaining to another subject (female). In this utterance, the subject in her first correction of the digit, shows many occurrences of the same digits: 'No, not NINE five nine, FIVE five nine.' The capitalization here is based on the authors' perceptual impression of prominence.

We may see, in the mandible height time function (third panel), that the first 'nine' is strongly uttered but the middle digit ('five') is markedly weak (the vowel is still heard clearly as a full diphthong /aJ/). The three digit is pronounced as an integrated single phrase. In contrast, the corrected digit sequence in the later part of the utterance is distinct, all three digits being separated by intervening silent periods (pauses). In particular, the last digit 'nine', which is not corrected, shows a large jaw excursion comparable to the corrected first digit. The last digit therefore shows a considerably expanded syllable duration, but the acoustic waveform (see bottom panel) shows marked amplitude reduction even in comparison with the middle digit which shows the smallest jaw excursion. We can thus appreciate the independence of observable variables such as the extent of jaw opening (or the amount of excursion), the amplitude of the acoustic signal (after subtracting the vowel identity effect), and the syllable duration which may vary depending on the definition significantly.

According to the C/D model, we interpret the length of each gap between contiguous syllable triangles to represent the magnitude of an inserted boundary. A triangle similar to syllable triangles could be created to exactly fill each gap by adjusting each boundary pulse height, assuming (asymmetric) fixed shadow angles also for boundaries. By doing so, we would be showing various boundaries in continuously variable and empirically determined magnitudes, reflecting the hierarchically phrased or bracketed structure of the linguistic form and numeric readjustment due to utterance effects.

Fig. III illustrates an utterance in a different dialogue by the same speaker as in Fig. II, correcting (the second time) the final digit saying 'five nine FIVE, the last number's a FIVE'. As shown in the top panel, there is a clear separation of the third emphasized digit 'FIVE' from the first two digits 'five nine', which is integrated and uttered relatively weakly. It should be noted that in this example in terms of jaw articulation, the first word is the weakest, the jaw opening gradually increasing toward the emphasized final syllable. However, the acoustic signal, as shown in the bottom panel, shows that the first syllable has an amplitude comparable to the emphasized last digit, while the middle digit ('nine' with the same vowel) appears distinctly weaker than the first.

Junping Gong in his preliminary study of mutual correlation among jaw opening, syllable duration as determined by acoustic discontinuities, and rms. of acoustic signal amplitude, demonstrated a striking lack of statistical correlation particularly between rms. and jaw excursion using a read-speech version of the Pine-street data. The current study as exemplified in this figure supports such a lack of consistent correlation even within the same utterance between articulatory and acoustic characteristics. If this example of utterance were typical, one might consider jaw opening and associated durational expansion to be more indicative of emphasis than acoustic signal amplitude, since in this case, the perceptual effect is also clear, indicating an emphasis placed on the third digit.

Fig. III also shows that, when the last of the three digits in a sequence was given prominence by the intended correction, this digit was preceded by a fairly large boundary (pause), roughly 100 msec. This computed boundary gap is about the same as the duration of the silent period as found in the acoustic signal in this case. Thin vertical lines in our figures are drawn to show where the ends of the syllable triangles are located in time.

4. CONCLUSION

The new method of syllable-boundary analysis seems to reveal much information about the prosodic organization of utterances reflecting some emotion of the speaker such as irritation as well as contrastive emphasis [8, 9].

There are many ways in daily conversation for correcting the dialogue partner's error using different linguistic forms, often accompanying some emotional elements. In any of such utterances, the (current version of) C/D model claims that the prosodic pattern can be represented simply by the syllable-boundary pulse trains with associated phonological feature specifications and numerical specifications of phonetic system parameters, local and global.

In the current study, the temporal changes have been interpreted to involve two distinct phonetic mechanisms: (1) syllable duration modulation, ascribed to syllable magnitude control and (2) insertion and modulation of boundary pulses. It seems that, while alterations of metrical structure may well take place at the level of phonological representation, the phonetic alteration of boundary strengths is inevitably continuous. The syllable/boundary pulse train representation may represent such information comprehensively. Whether the boundary in question is realized as an acoustic silence interval or just temporal adjustment such as elongation for the additional duration seems to depend on still unknown factors.

5. ACKNOWLEDGMENT

This work was supported in part by a NSF research grant (PIs: O. Fujimura and D. Erickson), and in part by support from ATR/MIC, Kyoto, Japan.

REFERENCES

1. Fujimura, O. "C/D Model: A Computational Model of Phonetic Implementation", in *Language Computations,* Ristad, E. (*ed.*), 1-20, RI: Am. Math. Soc., 1994

2. Fujimura, O. Syllable features and temporal structure of speech, in B. Palek (*ed.*), *Proc. LP'96,* 91-111, Prague: Charles Univ. Press, 1997..

3. Fujimura, O. Neuromuscular simulation and linguistic control. *Bulletin Com. Parlée,* in press.

4. Fujimura, O. "Syllable timing computation in the C/D model", in Proc. 3rd ICSLP, 1994.

5. Fujimura, O. "Syllables: Its internal structure and role in prosodic organization", in Palek, B. (*ed.*), *LP 94*, 53-93, Prague: Charles U. P., 1995.

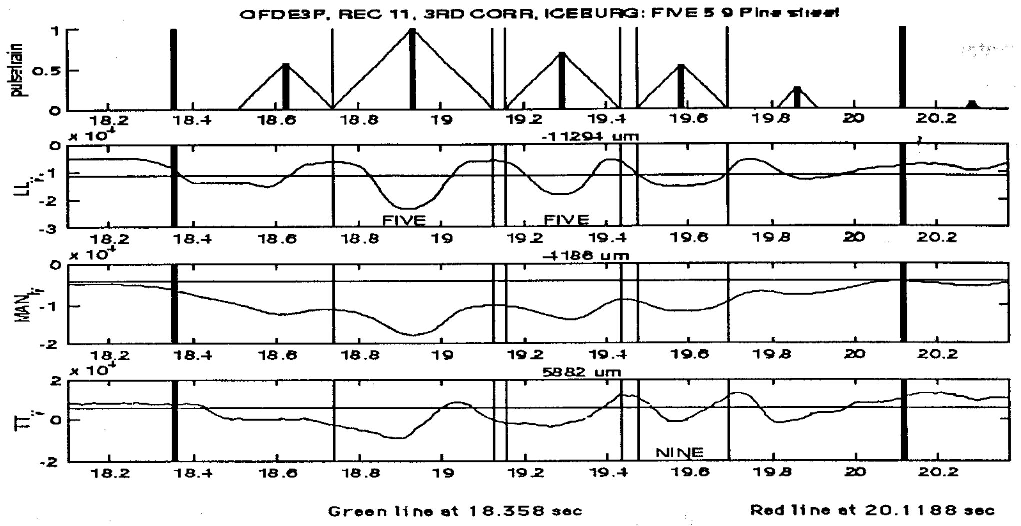
6. Erickson, D., Fujimura, O., Pardo, B. Articulatory Correlates of Prosodic Control: Emotion and Emphasis, in press.

7. Fujimura, O. "Relative Invariance of Articulatory Movements", in Perkell, J. S. & Klatt, D. H. (*eds.*), *Invariance and Variability in Speech Processes*, 226-42, Hillsdale, NJ: Lawrence Erlbaum, 1986.

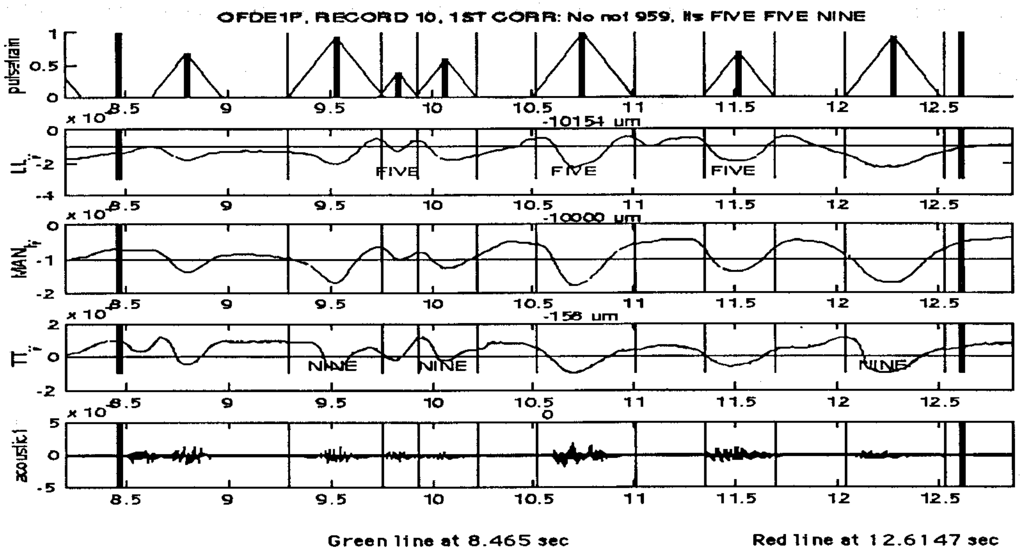
8. Erickson. Effects of Contrastive Emphasis on Jaw Opening. *Phonetica,* in press..

9. Spring, C., Erickson, D., and Call, T. Emotional modalities and intonation in spoken language. *Proceedings ICSLP92*, 679-682, 1992.

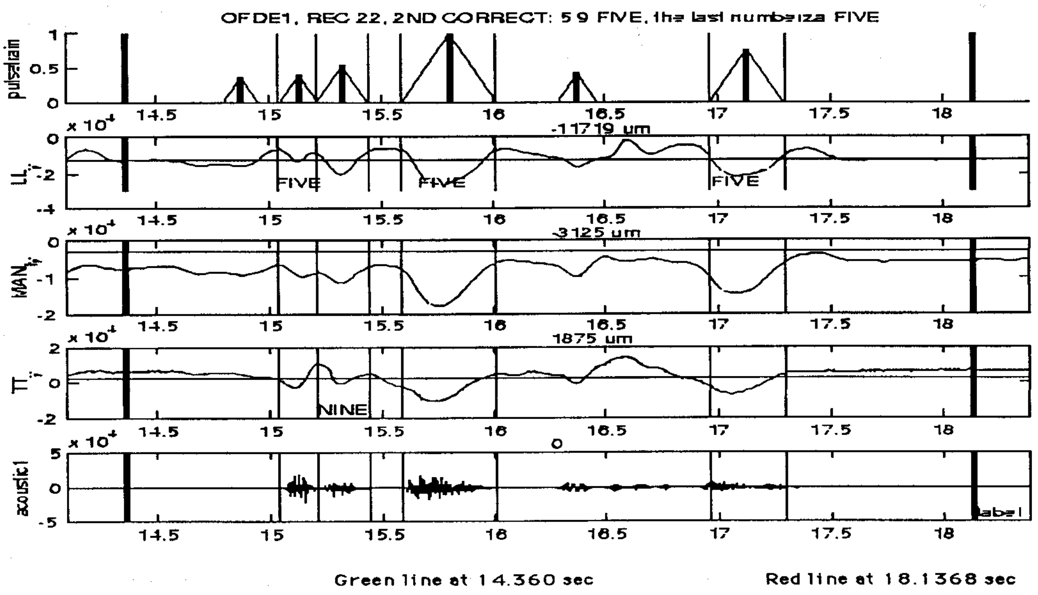
I


OFDE3P, REC 11, 3RD CORR, ICEBURG: FIVE 5 9 Pine street
pulsetrain
LL
MAN
TT
x 10⁴
-11294 um
-4186 um
5882 um
FIVE
FIVE
NINE
18.2
18.4
18.6
18.8
19
19.2
19.4
19.6
19.8
20
20.2
Green line at 18.358 sec
Red line at 20.1188 sec

II


OFDE1P, RECORD 10, 1ST CORR: No not 959, its FIVE FIVE NINE
pulsetrain
LL
MAN
TT
acoustic1
x 10⁴
-10154 um
-10000 um
-156 um
0
FIVE
FIVE
FIVE
NINE
NINE
NINE
8.5
9
9.5
10
10.5
11
11.5
12
12.5
Green line at 8.465 sec
Red line at 12.6147 sec

III


OFDE1, REC 22, 2ND CORRECT: 5 9 FIVE, the last numberiza FIVE
pulsetrain
LL
MAN
TT
acoustic1
x 10⁴
-11719 um
-3125 um
1875 um
0
FIVE
FIVE
FIVE
NINE
label
14.5
15
15.5
16
16.5
17
17.5
18
Green line at 14.360 sec
Red line at 18.1368 sec