CHANGES IN SYLLABLE AND BOUNDARY STRENGTHS DUE TO IRRITATION

Corey J. Mitchell[1], Caroline Menezes[1], J.C. Williams[1], Bryan Pardo[2], Donna Erickson[3], and Osamu Fujimura[1]

[1]The Ohio State University, Columbus, OH 43210-1002, USA
[2]University of Michigan, Ann Arbor, Michigan, USA.
[3] Gifu City Women's College, Gifu City, Japan

ABSTRACT

This study examines prosodic characteristics of speech in dialogue exchanges. Subjects were asked 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 obtained on four speakers of General American English at the University of Wisconsin x-ray microbeam facility. These data are analyzed using computational algorithms based on the theoretical framework of the Converter/Distributor (C/D) model [7]. The data analysis primarily pertains to jaw movements to evaluate syllable and boundary magnitudes in varied prosodic conditions, represented in the form of a linear syllable-boundary pulse train [7] that is interpreted as the rhythmic structure of an utterance. Preliminary results on syllable and boundary conditions indicate 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.

1. INTRODUCTION

Emotion can affect speech in terms of changes in temporal organization, intensity, stationary and dynamic characteristics of articulation (formants), and voice quality including F_0 and voice source spectrum envelopes [5], [13]. This paper addresses changes that occur in the rhythmic organization of an utterance due to a speaker's emotion, which may be labeled irritation, along with emphasis in conjunction with error corrections in conversational speech during dialogue exchanges between an experimenter and a subject [3]. The theoretical framework underlying the data analysis is the Converter/Distributor (C/D) model [7], which assumes that the prosodic organization of an utterance can be represented phonetically as a series of syllable and boundary pulses. These pulses vary in magnitude according to the abstract prosodic strengths (which are computable based on a phonetically augmented metrical structure), given as the input to the generative model of phonetic implementation. The C/D model deviates fundamentally from the classical segment concatenation and coarticulatory model (*e.g.,* [13], [6]). It uses syllables, rather than phonemes, as the minimal concatenative units, representing each syllable's internal structure directly in terms of phonological features and phonetic gestures. It represents speech signals as an organization of articulatory events, by a computational process of phonetic implementation. The input specification assumed for this mapping, from an abstract representation of an utterance to concrete articulatory or acoustic signals, includes specifications of various utterance conditions as system parameters, as well as paraphonological, discourse-related specifications for local prominence control supplementing the phonological representation of a linguistic form. Among other current models of speech production, Articulatory Phonology (AP) [1] adopts articulatory gestures at the lexical specification level. As mentioned by Krakow, AP seems not to have discussed any mechanism for representing articulatory variation that is not lexically specified [10], while C/D attempts a formal representation of utterance characteristics, including emotional expressions, as a comprehensive model of conversational speech [7].

Implementation of tonal specifications for lexical and phrasal accent features are incorporated in C/D into what is called the base function, along with vowel-to-vowel articulatory flow of movement, switching between voiced and unvoiced, prosodic mandibular control, and respiratory-phonatory control. The syllable-boundary pulse train represents the skeletal rhythmic structure of the utterance. Previous studies based on the theoretical framework of C/D have shown that syllable magnitude, as reflected by the amount of jaw opening, increases when the speaker becomes more irritated as well as with repeated corrections [14], [3]. These studies suggest that listeners tend to perceive more irritation as the number of corrections increase in the dialogue (see Dialogue Set 17 below). The current study describes a newly developed analysis method, which expands and refines previous measures [9], to infer the abstract syllable and boundary magnitudes from articulatory movement patterns, independently from acoustic signal characteristics.

2. EXPERIMENTAL METHOD

2.1 Data Acquisition

The data in this study were selected from a larger body of data collected in a previous study on the prosodic organization of conversational utterances [3]. Articulatory data with acoustic signals were recorded from 4 native speakers (2 male and 2 female) of Midwestern American English (Table 1). For this pre-print, however, data analysis was limited to speaker 3 (male). Analysis of irritation was limited to 3 corrections per

dialogue because it appeared that speakers did not show strong irritation beyond the third correction, perhaps due to resignation in later exchanges.

Speaker	Dialogues	Exchanges	Fives	Nines
1	11	41	66	57
2	10	53	78	56
3	14	72	123	93
4	16	79	138	99
Total	56	258	441	333

Table 1: Data corpus.

The experimenter (Donna Erickson) conducted a dialogue with the subject, whose articulation and acoustic signals were recorded, according to the general protocol of the x-ray microbeam facilities at the University of Wisconsin [15]. Spherical gold pellets (2.5-3 mm in diameter) were affixed to selected points on the speaker's tongue, lips, and jaw.

The target phrase in the dialogue always contained one of the following three-digit sequences: '595,' '959,' or '559.' The dialogue always included a reference utterance (the first of the dialogue) by the subject and used 'five,' 'nine,' or 'pine' as the target words in a street address. It was flexibly designed to allow the subject to use varied expressions in order to convey the message of a given street address. A typical dialogue is given below. In this dialogue, the speaker responded to the elicitor's "misunderstanding" of the final digit, noted with capital letters.

Dialogue Set 17, Speaker 2 (male)
Reference: DE: Where do you live?
S2: I live at 559 Pine Street.
1.DE: I'm sorry, that was 555 Pine Street?
S2: No, 55NINE Pine Street.
2.DE: I'm not hearing you, is it 555 Pine Street?
S2: No, it's 55NINE Pine Street.
3.DE: You're saying 555 Pine Street?
S2: No, there's a nine at the end.
It's 55NINE Pine Street.
4.DE: 555 Pine Street, right?
S2: No, 55NINE Pine Street.

The same vowel, /aJ/ (a palatalized diphthong), was used for all the key words to observe prosodic effects free from vowel effects. In a previous study [3], jaw position was found to change significantly for emphasized *vs*. unemphasized key words with the vowel /aJ/.

2.2 Automatic Inference of Syllable Triangles

All articulatory data from the Pine Street speech production database used in the present study were available for analysis on an IBM-compatible computer using a special data interpretation program, UBEDIT, created on the MATLAB platform by Bryan Pardo. In this study, mandible height, lower lip height, and tongue tip height were selected for our display, along with the acoustic waveform (Fig. 2). UBEDIT evaluates and records the occurrence in time of the minimum mandible height (maximum jaw opening) for each target syllable. A pulse is then erected (top panel): the magnitude of the pulse is the distance between the occlusal plane and the jaw minimum position and it is tentatively placed in time at the point when the minimum occurred.

2.3 Readjustment of Syllable Timing

In order to more accurately estimate the timing of the syllable pulses, a version of the "iceberg" method of movement timing evaluation [9] is used. For each syllable considered in this study, there is a single articulator which implements the place of articulation for onset or coda (*i.e*., the lower lip for 'five' and the tongue tip for 'nine'). This is the critical articulator for the demisyllable. For each critical articulator, a vertical "iceberg threshold" was determined as follows. Given the set of *n* instances of a given syllable uttered by a given speaker (for example, 57 examples of speaker one saying 'nine'), the time function representing vertical position, sampled each 6.9 msec. for the critical articulator, was determined. One millimeter wide bands (contiguous and non-overlapping) were created and all data points (relating velocity to position) from all utterances of the syllable were placed in a band defined by their position. An "iceberg metric" was computed for all bands in the range of motion of the critical articulator, according to the formula:

$$M(i) = (C * AvgVelVar(i\text{-}1 \text{ to } i + 1)) / \text{abs}(MeanVel(i)),$$

where i is the index number of each position band, and C is a constant scaling factor. $AvgVelVar(i\text{-}1 \text{ to } i + 1)$ is the variance in absolute value of the velocity of all data points in bands i-1, i, i+1, and $\text{abs}(MeanVel(i))$ is the absolute value of the mean velocity of all data points found in the i-th band. Then the band with the smallest value for M is selected to give the iceberg threshold height.

The midpoint between the threshold crossing time of the descending demisyllabic time function and that of the ascending demisyllable yields the "center" of the syllable, to which the time position of the syllable pulse is readjusted. A scatter plot (velocity against position), showing the threshold setting value (circled point on the graph) for the digit 'five,' for speaker 3, is shown in Figure 1. In this computation, position has been truncated to 1mm increments. The threshold (horizontal pellet height) line for each crucial articulator is then set at a constant y value, for each dialogue of a speaker.

2.4 Inference of Articulatory Syllable Boundary and Durations

The program then automatically constructs a symmetric triangle with a maximum "shadow" angle, drawing slanted lines from the top of the syllable pulse to the horizontal line at its foot (on both sides), without allowing any overlap throughout the dialogue. This angle is adopted as the critical angle for all the syllable triangles of target words ('five,' 'nine' and Pine') in the given dialogue. The program then automatically records the beginning and ending time values for each syllable triangle and evaluates "syllable duration" as the temporal distance between the two edges. If there is any gap between consecutive syllable triangles, this is interpreted

to represent a boundary, and the gap length in time is interpreted as the boundary duration, proportional to an abstract measure of boundary magnitude. Often, however, it has been found that for each critical shadow angle thus determined for a dialogue, many contiguous syllable triangles leave no significant gaps, in which case we interpret that there is no phonetic boundary between the syllables.

3. RESULTS AND DISCUSSION

3.1 Acoustic Duration, Emphasis, and Irritation

In a previous study using the same database [4], acoustic durations for the digits 'five' and 'nine' were measured using the waveform and spectrograms on WAVES+ESPS software. The digit 'five' was measured from the onset of the frication constriction for /f/, where there is an audible coming together of lips for initial /f/. The offset of /v/ at the point of cessation of voicing was taken for the end of acoustic duration for the digit 'five'. The digit 'nine' was measured from the onset of the initial nasal murmur to the offset of the coda nasal murmur. These measurements were used for evaluating the correlation between acoustic duration and irritation in this study.

Some related work has been done in which those digits judged to be emphasized by listeners were examined in terms of acoustic durations [4]. In the current paper, we analyzed the acoustic duration patterns in the same database measured by [4].

Correction	Initial	Middle	Final
No C (n=14)	411 (44)	344 (39)	431 (37)
C digit 1 (n=14)	381 (58)	385 (42)	391 (60)
C digit 2 (n=18)	369 (68)	418 (49)	394 (57)
C digit 3 (n=13)	366 (63)	334 (63)	461 (51)
Total (n=59)	381 (60)	374 (59)	417 (58)

Table 2: Means (S.D.) for acoustic duration and corrected digit in msec.

Table 2 shows the means for acoustic duration for each of the three digits corresponding to intended emphasis (*i.e.*, the corrected digit). Different rows are for utterances that contain correction of different digits. For example, the row with C digit 1 reports that there were 14 utterances that had correction of digit 1 for which the duration of the initial digit had a mean of 411msec. (S.D. 44 msec) Generally, the corrected digits (regardless of perceived judgement of emphasis) are longer in mean duration than the corresponding digits in the reference utterances (no C) by the same speaker. The only exception was the initial digit which was shorter, when corrected, than in reference utterances. In addition, there seems to be a tendency of digits other than the corrected digits to be shorter in duration in comparison with the corresponding digits in the reference utterances. The middle digit, however, does not show this tendency.

Irritation	Initial	Medial	Final
0 (n=11)	385 (59)	345 (63)	445 (64)
1 (n=19)	386 (71)	369 (67)	402 (54)
2 (n=10)	366 (55)	358 (54)	416 (63)
3 (n=7)	360 (61)	391 (40)	381 (35)
4 (n=8)	385 (67)	405 (60)	416 (60)
5 (n=7)	390 (58)	395 (84)	425 (31)
6 (n=3)	361 (111)	327 (41)	394 (41)
7 (n=3)	351 (43)	389 (46)	363 (53)
8 (n=6)	367 (77)	392 (90)	430 (74)
Total(n=74)	377 (64)	373 (65)	412 (57)

Table 3: Means (S.D.) for acoustic duration and irritation in msec.

Table 3 shows the means for acoustic duration for perceived irritation. Perceived irritation scores in this report are the number of listeners who selected "irritated" from five emotional states when the whole utterance of the digit sequence was presented [3]. For example, the second line of this table shows that 19 of the 74 utterances of speaker 3, were given the label "irritated" by 1 of the 8 listeners. In those utterances, the mean of the acoustic duration for the initial digit was 386 ms (S.D. 71), for the middle digit 369 ms (S.D. 67), and for the final digit, 402 ms (S.D. 54). These duration values are calculated for all conditions regardless of corrections. These data are based on the acoustic duration measurements reported in [4].

This table does not seem to show any clear pattern of correlation between the acoustic duration and perceptual measure of irritation in the form of the numbers of listeners who selected the label "irritated" for the utterance that contained the digit in question. Correlation between irritation and acoustic duration yielded an $r = .422$ ($p < .01$). Only 18% of the irritation could be explained by acoustic duration ($r^2 = .178$). However, with irritation as the dependent variable and acoustic duration and intended correction as the independent variables, the increase in correlation revealed considerable interaction among the variables ($r = .656$ ($p < .015$).

Incidentally, it was found that as the correction was repeated within the dialogue, there was a tendency for the acoustic duration of the corrected digit to increase, indicating that the emphasis on the corrected digit was reinforced as the correction was repeated more times.

3.2 Syllable Duration, Emphasis, and Irritation due to Correction

Syllable duration, as computed based on jaw opening, correlated significantly with irritation with an $r = .550$ ($p < .0001$), that is, 30% of the irritation score is explainable by syllable duration. When corrected digit and syllable duration were correlated with irritation, $r = .611$ ($p < .001$) and $r^2 = .373$ (37%). The correlation score increased significantly ($r = .667$, at $p < .004$, and $r^2 = .444$ (44%) when corrected digit and syllable duration were considered, revealing considerable interaction.

Table 4 shows the means of syllable duration for different scores of perceived irritation. From these tables it is evident that there is an increase in syllable duration as irritation increases for middle and final digits. Syllable duration and corrected digit together correlated significantly, r= .838 (p< .001) with the number of corrections the subject made within the dialogue (70%).

Irritation	Initial	Middle	Final
0 (n=8)	283 (52)	252 (38)	223 (46)
1 (n=19)	328 (73)	284 (67)	264 (47)
2 (n=10)	294 (65)	244 (60)	226 (47)
3 (n=7)	352 (22)	319 (40)	281 (28)
4 (n=6)	316 (84)	307 (82)	255 (50)
5 (n=7)	304 (68)	276 (68)	257 (45)
6 (n=2)	378 (2)	326 (50)	283 (17)
7 (n=1)	303	330	266
8 (n=6)	331 (32)	309 (48)	297 (23)
Total(n=66)	318 (63)	284 (62)	258 (47)

Table 4: Means (S.D.) for syllable duration and irritation in msec.

3.3 Boundary Duration, Emphasis, and Irritation due to Correction

Fig 3 and Fig. 4 show the mean boundary duration for different conditions of digit correction. In fig. 3 the boundary strength between digit 1 and 2 is significantly larger when it follows the corrected digit (C on digit 1). In fig 4 the boundary strength between digit 2 and 3 is significantly larger when it precedes the emphasized digit (*i.e.,* the correction on digit 3). Spring *et al.* reported that contrastive emphasis (corrected digit) was better perceived for the initial digit when it was corrected, and not so well perceived for the final digit [14]. At the same time emotion was better perceived for a digit sequence when the utterance's final digit was corrected rather than its initial digit. The correlation of boundary duration between digit 1 and 2 in general with irritation (Fig. 5) was not significant, r = .072. The correlation of boundary between digit 2 and 3 with irritation (Fig. 6) was significant r = .237 (p< .08). Five percent of perceived irritation was accounted for by the boundary between digit 2 and 3. The correlation of both boundary durations together with irritation was significant at .063, r = .317. When both boundary durations and all syllable durations were considered as independent variables, with irritation score as the dependent variable, the correlation coefficient of r = .576, that is 33% of the irritation could be explained by boundary duration and syllable duration.

In general, irritation increases with syllable duration positively. When boundary strength is added into the equation, the correlation is increased. When boundary duration, syllable duration, and corrected digit were regressed on irritation, the correlation was r = .787 (p<. 002), *i.e.*, boundary strength, syllable strength and corrected digit can explain 62% of perceived irritation. Boundary duration and corrected digit together correlated significantly, r = .784 (p< .001) with number of corrections the subject made.

The significant increase in boundary duration between digit 1 and 2 when the first digit was corrected could possibly explain why subjects were better at perceiving corrected digits in this condition. The long boundary duration between digit 2 and 3 before the final digit could be the reason subjects perceived emotion rather than digit correction in this context. Whether it is generally true that preceding strong boundaries are associated with irritation and succeeding strong boundaries with digit correction is a subject of further research.

4. CONCLUSION

The results of this paper suggest that changes occurred in the rhythmic organization of an utterance due to a speaker's emotion (*i.e.*, irritation), digit correction, and number of repetitions of the same correction the speaker had to make during the dialogue exchange. Changes were observed in syllable duration and boundary duration. Generally, when a syllable was corrected, there was significant increase in the duration of the syllable. Moreover, as the number of repeated corrections increased, the syllable duration also increased.

Boundary durations changed significantly when digits were corrected. The boundary between the first and second digits was significantly larger when correction was on the first digit. However, boundaries between digits 2 and 3 were larger when correction was on the final digit. The results on boundary duration as reported by this paper indicate a possible change in rhythmic organization which listeners use as a cue for distinguishing corrections and emotion (irritation). The new algorithm for inferring syllable durations, and thereby boundary durations, seems to reveal interesting temporal reorganization patterns of prosodically variable utterances.

5. ACKNOWLEDGEMENT

This research has been supported in part by NSF (BCS-9977018), NSF (SBR-951198) and ATR, Japan.

6. REFERENCES

1. Browman, C. P. and Goldstein, L. M. Towards an articulatory phonology. In C. Ewan and J. Anderson (Eds) *Phonology Yearbook 3*. Cambridge University Press, Cambridge, UK, 219-253, 1986

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

3. Erickson, D., Fujimura, O., and Pardo, B. Articulatory correlates of prosodic control: emotion and emphasis. *Language and Speech 41*, 395-413, 1998

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

5. Fant, G. *et al.* The source-filter frame of prominence. *Phonetica*, in press

6. Farnetani, E. Coarticulation and connected speech processes. In W. Hardcastle and J. Laver (Eds) *The handbook of phonetics*. Cambridge, MA Blackwell Publishers 371-404, 1997

7. Fujimura, O. The C/D model and the prosodic modulation of articulatory behavior. *Phonetica,* in press

8. Fujimura, O. Relative invariance of articulatory movement. In J. Perkell, and D. Klatt (Eds) *Invariance and variability in speech processes*. Lawrence Erlbaum, Hillsdale, NJ, 226-242, 1986

9. Fujimura, O., Pardo, B., and Erickson, D. Effect of emphasis and irritation on jaw opening. *Proc. ESCA*, 23-29, 1998

10. Krakow, R. A. Physiological organization of syllables: a review. *Journal of Phonetics, 27*, 23-54, 1999

11. Laver, J. *Principles of phonetics.* Cambridge University Press, Cambridge, UK, 1994

12. Levelt, W.J.M. *Speaking: from intention to articulation*. MIT Press, Cambridge, MA, 1989

13. Lindblom, B. Spectrographic study of vowel reduction. *JASA 35*, 1773-1781, 1963

14. Spring, C., Erickson, D., and Call, T. Emotional modalities and intonation in spoken language. In J. Ohala, *et al.* (Eds) *Proc. ICSLP-92*. University of Alberta, Canada, 679-682, 1992

15. Westbury, J. The significance and measurement of head position during speech production experiments using the x-ray microbeam system. *JASA 85*, Suppl.1, S98, 1994

7. FIGURES

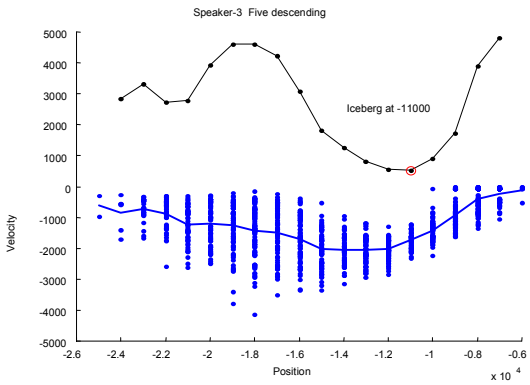

Figure 1: Determination of iceberg threshold height (circle). Position is in μm (10^4).

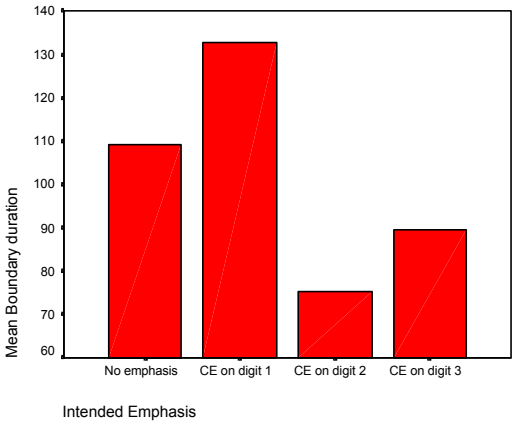

Figure 3: Mean boundary duration (msec.) between digits 1 and 2 as a function of corrected digit.

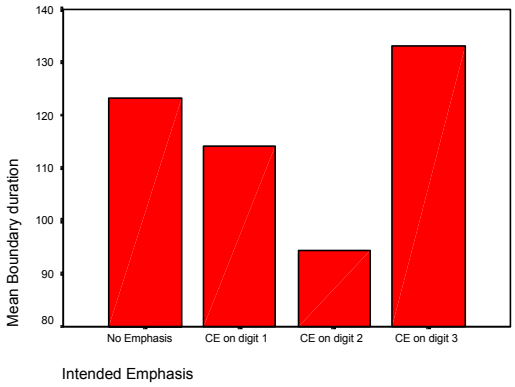

Figure 4: Mean boundary duration (msec.) between digits 2 and 3 as a function of corrected digit.

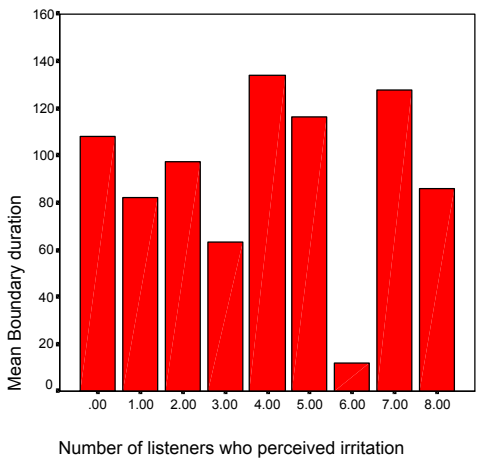

Figure 5: Mean boundary duration (msec.) between digits 1 and 2 as a function of perceived irritation.

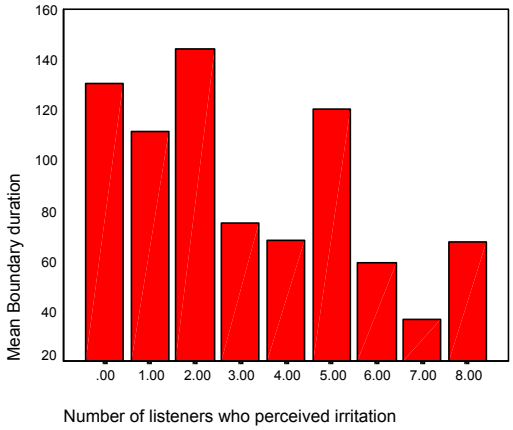

Figure 6: Mean boundary duration (msec.) between digits 2 and 3 as a function of perceived irritation.

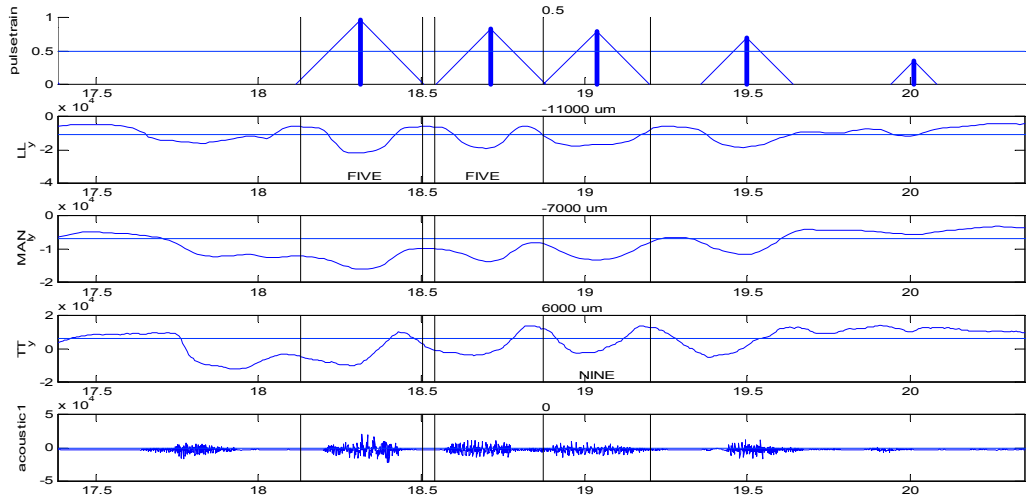

Figure 2: A computer screen display showing, on the top panel, syllable triangles for Five, Five, Nine, Pine and Street from left to right. X-axis is in sec.; y-axis is in µm ($x10^4$).