ON THE ROLE OF DURATION IN CONNECTED SPEECH

Antônio Roberto Monteiro Simões[*]

Abstract: In this study it is argued that the main role of duration in connected speech is to direct attention to statistically significant variations in sound segment lengthening and reduction, during speech acts. If this role is confirmed in actual discourse, then we can expect it to permeate all linguistic components, from Phonetics, to Phonology, to Syntax, Semantics and Pragmatics. This conclusion is supported by data obtained in Simões' (1987) study of duration in speech synthesis by rule.

Keywords: Duration. Model. Synthesis by Rule. Brazilian Portuguese. English. Phonetics. Klatt.

Resumo: Neste estudo argumenta-se que o papel principal da duração na fala conectada é o de assinalar as variações estatisticamente significantes, em termos de alongamento e redução de fones, durante os atos de fala. Se esta função for confirmada no discurso, neste caso pode-se esperar que este papel específico da duração permeie todos os componentes linguísticos, desde a fonética à fonologia, à sintaxe, à semântica e à pragmática. Esta conclusão baseia-se em dados obtidos no estudo de Simões (1987), sobre a duração em voz sintética produzida por regras.

Palavras-chave: Duração. Model. Síntese Por Regra. Português do Brasil. Inglês, Fonética. Klatt.

Introduction

The role of sound segment duration in discourse often intrigues me, given my skepticism with the common explanations of its role in discourse. To me, the statistically significant variations of duration in terms of shortening and lengthening of sound segment serve to point out significant speech events in spontaneous discourse. A statistically significant difference in duration at certain points in spontaneous speech is a signal or a flag of a significant speech event, at the points where these variations occur. Examples of statistically significant events come from all linguistics domains, e.g. sound segment

[*] University of Kansas, Lawrence, Kansas, USA, asimoes@ku.ed

lengthening of vowels or consonants, sound segment reduction of vowels or consonants, word shortening or lengthening, shortening or lengthening of phrase or sentence beginning or ending. For instance, a word lengthening can be the result of the event *word novelty*. Therefore, speech events happen in all sentence tiers. Variations in duration will point where these events happens.

The consequence of the claim made in this analysis is that it changes the focus from seeing duration as a parameter per se into mainly an indicator of speech events. Such a view does not minimize the importance of duration in speech. Duration will always deserve a great attention. As stated in Fant (1970) "The simple and fundamental cue of duration deserves greater attention than is conventionally paid to it."

In order to support the view of this study, data from Simões (1987) is revisited and reanalyzed. In Simões's (1987), an adaptation of Klatt's (1976) model for English was developed for Brazilian Portuguese. Klatt's (1976) model was intended to create speech synthesis by rule in English. In other words, Klatt's study was intended for artificial speech, but it was based on data obtained with natural languages. Although it is obvious that synthetic speech is not necessarily a mirror of natural mental processes, in the case of synthesis by rule, it may provide us with insights into natural language processes.

Experimental protocol

The experimental protocol in Simões (1987) was organized according to three main procedures: the production of the corpus (recordings), the production of spectrograms for sound-segment segmentation, and data analysis.

The corpus for analysis was obtained from recordings of a native speaker of Brazilian Portuguese, PM, 32 years old, from Rio de Janeiro, was recorded. PM read a 1286-word text for children, titled *O sonho de Carolina* written by Vera Campos Ferrão, in two recording sessions. The recording sessions took place one week apart. PM was asked to read the same text three times in each session, totaling six readings of the same story. The third reading of the second session, i.e. the sixth recording, is the one used in the analysis.

The recordings were made in 1984, at the UT language laboratory, in Batts Hall with the help of a lab technician. An acoustically isolated recording booth was used and the recording was done using a AKG dynamic, unidirectional cardiod microphone situated at 40 cm from the subject's mouth. Before each recording the system was calibrated. The tape used

is an AMPEX tape, 1/4", 1.5 mil, mylar.

The segmentation of sound segments was realized as follows. Some 700 spectrograms were produced to observe and measure the 1286-word text and its sound-segments in this study. The spectrograms were obtained through a stereo tape deck TEAC A-2300SX coupled to a Digital Kay Sonagraph 7800. In the eighties, the analysis of connected speech required a special use of the Sona-Graph. To use only the usual 3D broad band spectrograms would require too many arbitrary decisions. The technique I developed combined two spectrograms for each sample analyzed. One spectrogram had the regular broad band and a second one was made using the amplitude contour. Superimposed on the top of each spectrogram there was an oscillographic image. Extensive work done by the author using continuous speech had shown that observations and measurements of sound-segments in continuous speech were made reliably if this technique were used.

For a clearer image, a high shape range (AGC) was used in making the spectrograms. High shape range prevented amplitude from lowering in high frequencies. One then, had to be careful when looking at these figures in the high frequency such as the region of F3 and above. In other words, the use of AGC created artifacts which highlighted information which should be fading in that area. In this study observations were made at first, based on changes at Fl and F2 domains.

The seminal work of Delattre et al (1955) at Haskins Laboratory used a set of consonants [b,d,g] in contact with various vowels and found that the formant transitions of Fl and F2 are cues that caused the listener to identify these consonants. In their study, and this claim is still valid, it was observed that the formant transition of F2 seemed the most relevant cue to the understanding of these sounds. For this reason the discussions in the present analysis were mostly based on observations done at the first and second formants domains.

Some of the procedures in Simões (1987) had been used in acoustic analyses of speech in different laboratories by word of mouth. However, when the current rules were developed, it seems that it was the first time that specific rules of segmentation had been written. On the other hand, the measurements made with superimposition of different spectrographic views were a result of techniques of Simões' (1987) investigation.

Segmentation of speech-sounds in connected speech can be deceiving, and two people segmenting the same document will not show exactly the same measurements. The solution to this problem is to have the same person or the same program to make the measurements. Therefore, in order to create statistically valid data, the same researcher must take all

measurements, using rules for segmentation that are established and kept rigorously.

Klatt's (1975) way of segmenting served as the base for the segmentation procedures or rules in Simões (1987). Therefore, the notions of onglides, offglides, steady state, and simple and complex nuclei found in the literature (e.g. Klatt (1975), Lehiste and Peterson (1961)) were used.

Therefore, the manual segmentations were established according to the type of sounds under study, namely, the vowels. The visual cues which allowed reliable segmentation using the Digital Kay Sona-Graph are listed below in order of importance. When one procedure did not suffice, we would go to the next or combine two or three procedures. As mentioned above the Digital Kay Sona Graph can combine an oscillographic image with the usual 3D image which eliminated many problems one has in segmenting. Before any of the rules for manual segmentation below is applied these spectrogram pairings had to be well analyzed qualitatively. This preliminary analysis will allow for a first attempt in locating phonetically realized speech-sounds. Only then, are the rules below to be applied.

Rules for manual segmentation:

Rule 1. The point of departure is the onglide and the offglide **(LEHISTE and PETERSON** 1961) of the second formant transition, i.e. when the second formant starts (most of the times preceded by a *blank*), and when it is interrupted. In other words, the whole formant transitions are included as part of the sound to which they belong. In case there is a burst or a glottal stop, the segmentation is done before the burst or glottal stop, still at the F2 level.

Rule 2. Use the waveform information provided by a superimposed oscillograph built in the Digital Sona-Graph. This sound wave image shows clear variations when the glottal pulses (vertical spikes) shorten vertically to become almost confused with the zero lines. Such variations are indicative of the end of a sound-segment and/or the beginning of another. This is a consequence of damped oscillations. Any other sudden change in the amplitude of these soundwave images are potential indicators of sound boundaries, although these sudden changes are observed outside the zero line region.

Rule 3. Observe the changes in the relative intensity in the first formant region, reflected in the darkening of images. Any change in energy concentration might indicate a new sound. It is known that true consonants have less energy than true vowels. In terms of sound relative intensity (reflected in the darkening differences on the spectrogram), consonants have less intensity than vowels. This may be explained by the production of both classes of sounds. Vowels find no obstacle in their way out of the vocal tract and are realized with most of its energy from the glottal source. On the other hand, consonants are completely or partially obstructed in their realization creating a loss of energy.

Take into account that there are differences in the relative intensity of speech-sounds in the region of the first formants. Consonants for instance, are characterized by weaker energy (less darkening in the spectrograms) than vowels.

Rule 4. Take into consideration the lowering or rising of the fundamental frequency observed through spikes spacement (glottal pulses). Lowering of the Fo (fundamental frequency) happens when, relative to the glottal pulses of the preceding adjacent sound-segment, the glottal pulses are more apart; during rising of the Fo, glottal pulses are closer, relative to the glottal pulses of the preceding adjacent sound-segment.

Rule 5. As Klatt (1975) suggested, the burst characteristic of stop consonants, is considered as part of the following sound-segment, not as part of the stop. This is helpful in treating vowel segmentation when vowels are not preceded by stop consonants. The spectrograms in this study show that a glottal stop (visually similar to a burst) precedes all vowels most of the times, especially the front low [a]. In addition, liquids are also preceded by a glottal stop. Finally, any aspiration, including that associated with palatalization is included in the vowel portion.

The rules above should suffice for manual segmentation. However, in case there are doubts, additional information can be used. For instance, in the high frequency regions, namely regions above 4 kHz, there is usually a clear distinction between vowels and consonants. The formant structure of vowels may show four formants. The fourth formant gives information about the limits between two adjacent vowels in different words, when they happen to be apart. Also the anti-resonance characteristic of nasals can sometimes be used as a cue in separating the vowel portion from adjacent nasals.

Data analysis and discussion

This study has not exhausted the number of factors influencing duration patterns in speech, such physiological, discourse, and extra-linguistic factors, to mention some, which are all to be considered in dynamic approaches. This study covers some of the possible factors or domains, i.e. (1) the phonological and the phonetic domain, (2) the word domain, (3) the syntactic domain and (4) the semantic domain. As one can see multiple factors operate in reducing and expanding sound-segment duration.

As an illustration of the many factors influencing duration, we can look at one of the passages read for this study, "Carolina, a minhoca…" (Translation: *Carolina, the worm*…) In this passage, the last "a" in "Carolina" was not physically or phonetically realized. Some of the factors that could have caused the physical deletion of the "a" are

phonology: posttonic final position (word domain), unstressed; *rhythm, immediately after* a rhythmic foot; *intonation,* the [a] is in the minimum of a curve.

phonetics: postonic vowel having phonetic identity with the following vowel, due to a linking caused by the lack of an expected pause.

The statistical analyses were made using the S Statistical Package, implemented in the UNIX operating system (Unix is a trademark of Bell Laboratories), and run on a Digital Equipment Corporation VAX 11/780 at the UT Austin campus was used for most of processing and graphics. The analysis of variance used here is a local program (ANOVA8) which is based on the PMDP series. It was developed by Thain Marston, back then a graduate student in Experimental Psychology at UT Austin.

Analysis of all data from these vowels in the corpus consistently shows functions positively skewed, that is lower values concentrate on the left side of the horizontal axis (abscissa), in such a way that higher values will spread rightward. A parameter such as duration is expected to be positively skewed. This distribution is the reason the **median** was chosen as the measurement type for this parameter, instead of the mean. Observations of the behavior of these functions led to the definition of unmarked (*inherent*) duration. This is a definition which is quite objective and had not previously been proposed, before this study.

Former definitions of unmarked values are not disregarded. They are all valid and have already been proven useful. My proposition allows for researchers without access to various instruments to establish intrinsic values in an experimentally valid manner. Besides, the median gives an impartial decision which does not depend on previous knowledge of the speech under analysis. The use of the longest *inherent* duration (KLATT, 1976; LINDBLOM and al 1981), for example is not possible in CV languages since it requires the use of CVC monosyllables.

In Figure 1 a common pattern of the data is shown. Observe that the values of measurements of duration tend to cluster on the left side of the abscissa. Without considering yet factors such as stress and position, a first look into this curve can be done in any of the two approaches. One approach examines values falling after a given number of standard deviations (2 or 3 standard deviations). Values in these areas outside 2 or 3 standard deviations indicate sound segment durations that are statistically significant. The second approach analyzes values falling outside the quartiles of the distribution. The second approach is simpler because the statistical program used here lists the statistically significant data according to the quartiles leaving the user free from calculating standard deviations.

Figure 1: As expected from a parameter like duration, the data shows a pattern of distribution which is positively skewed. Quartiles and high values are indicated[1].

[1] This figure 1 and the next ones, Figures 2 and 3 are the original figures taken from Simões (1987).

Considering the approach using standard deviations, values falling on the left side of the median (the inherent duration) are difficult to handle because they cannot fall below the 0.0 value. They are thus biased because of the nature of the phenomenon duration. Some possibilities have to be considered in these cases. Maybe only one or one and a half standard deviations are to be taken as cutoff point, instead of two or three as it is done usually with a normal curve. This can bias the research since subjective decisions may take place. A more applicable solution would be to take a common logarithm of these positively skewed distributions and transform them into a normal distribution.

The values detected as statistically significant on the right side will serve as a reference point. Suppose the values 18.0 cm (144ms) are shown as statistically significant. Then the logarithm of 18.0 gives the cutoff line on the left side of the now normalized curve.

This allows for separating values which are simply expected values of these vowels from values which appear in the furthermost areas of the curve (*curve tail*). The values falling in the statistically important areas are the ones of interest and the ones that need explanations as to why are they falling far off the norm. It may be that a syntactic component is making

some vowels longer; it may come from melodic curves, from the phonological component, from the three of them and so on.

A step further in this approach combines general linguistic knowledge with experimental research, although extremely difficult to handle, should provide for a finer and more elegant suggestion as to what is happening in the speech chain. The next section concentrates on this aspect of the methodology used here.

Establishing stress groups

The linguistic contexts of the vowels analyzed in this study were organized as follows. Once the two classes of inherently stressed and unstressed words were properly assigned, then decisions were made about which syllables were pretonic (prl, pr2, pr3, ...) or postonic (postl, post2, post3, ...), according to the main stress (+stress). Therefore, considering, for example, a sentence like

"Carolina, a minhoca, tem dois grandes desgostos na vida -- ser careca e não ter cintura," (Carolina, the worn, has two big misfortunes -- to be bald and to lack a waistline.")

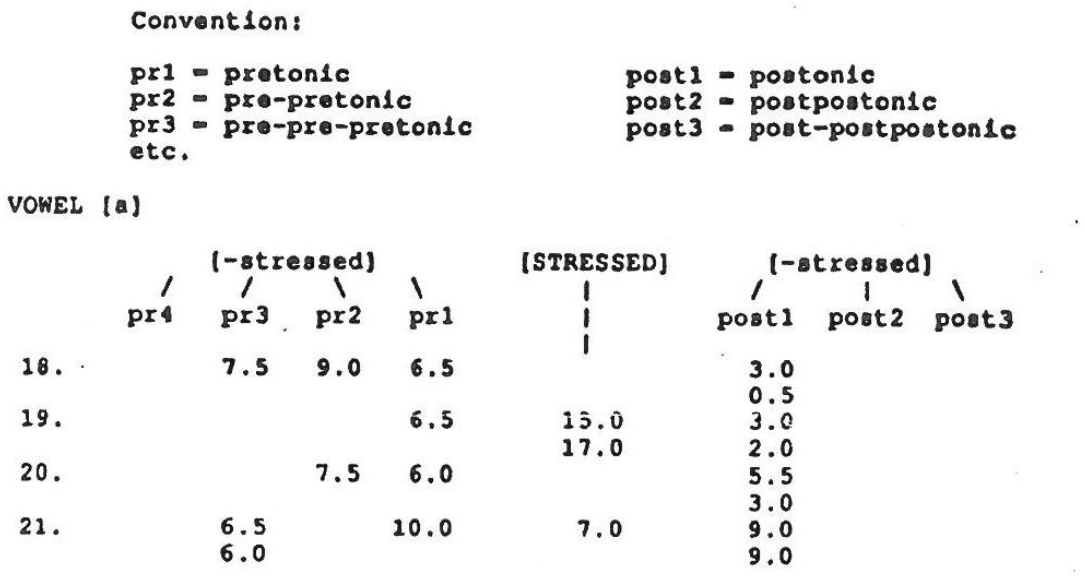
we obtain the breakdown shown in Figure 2:

Figure 2: Sentence breakdown in stressed and unstressed groups. This example uses the sentence "Carolina, a minhoca, tem dois grandes desgostos na vida -- ser careca e não ter cintura," (Carolina, the worn, has two big misfortunes -- to be bald and to lack a waistline.")

(...)pr3	pr2	pr1	+STRESS	post1	post2 (...)
	Ca	ro	li	na,	
	a	mi	nho	ca,	
				tem	
				dois	
				gran	des
			des	gos	tos
			na	vi	da
		ser	ca	re	ca
			e	não	
				ter	
			cin	tu	ra.

Once stressed and unstressed words were classified, syllables were classified according to their pretonicity and postonicity. Figure 3 illustrates this:

Figure 3: An example of how stressed and unstressed groups were classified inside the whole noun phrase.



Taking the opportunity of the great number of measurements taken for this corpus, these groupings were treated the same way above, namely the behavior of their functions was observed individually and then normalized with their common logarithm. Then other techniques of analysis could be used according to the goals. For example, analysis of variance was applied, in an attempt to find empirical support to the hypothesis that phonological processes operate at a given hierarchy, namely from postonic to both pretonics and to stressed positions.

The importance of sound duration in natural speech is a well-documented fact and it can be studied as far back as the first grammars were written. One of the first systematic studies of duration in our times can be traced back to Martinet (1949) who found that there existed a universal tendency in languages for tense consonants (/f,s,š,p,t,k/) to shorten vowels following them and for lax consonants (/v,z,ž,b,d,g/) to lengthen vowels following them. Jakobson et al (1952) observed that tense consonants are longer than lax consonants; Fry (1955) claimed that in his data duration and intensity ratios showed to be both cues for

judgments of stress, and that duration ratio is more effective than intensity ratio; similarly, Miller and Nicely (1955) proposed an acoustic-physiological feature *duration* to distinguish /s,š,z,ž/ from 12 other consonants. According to them these consonants were, long, intense, and contained high-frequency noise, but it was their extra duration which is most effective in setting them apart. Earlier studies of rhythmic patterns that dispute the role of duration in speech rhythm. For instance, Kozhevnikov and Chistovich (1965) proposed a speech production model using the average length of seven syllables as a rhythmic foot. In other words, the rhythmic program for a word is independent of the differences in the length of the word. Noteboom (1972) repeated these experiments in his analysis of Dutch and confirmed the idea that subjects are much more aware of duration if a monotonous pitch is used than if normal pitches are used. Intonation affects duration, but rhythmic program for a word is independent of the difference in *length* and *shortness* of words. On the other hand, well before these experiments, Pike's (1945) seminal study of English claimed that rhythm may be dependent of word duration.

The preceding paragraph is not intended to give an exhaustive list of works on duration, but simply to confirm the power of this parameter, and the great variety of studies based on duration. For a current review of studies on duration, LAZARIDIS et al (2012) offer an extensive list of studies in their article to improve the quality of prosody in speech synthesis from text.

In sum, the role of duration has been studied in relation to all linguistic domains, e.g. phonology, phonetics, prosody, word domain, syntax and semantics. Models of speech production take duration into account. Therefore, it is undeniable that duration permeates all aspects of speech production and perception. All linguistic phenomena are connected to time, to duration.

It is still premature to deny the main role of duration in setting apart phonemes. Meanwhile, a question that one may ask is if it is really possible to hit the correct target by simply lengthening a sound segment in natural languages.

In Italian, for instance, lengthening seems to be a result of tone inflections of the vowel adjacent to a so-called long consonant, e.g. the "u" before "tt" in *tutto* (all). With regards to consonants, it is more plausible to talk about geminates or repeated consonants, when a consonant repetition is clear, as the "mm" in the French word *sommet* (peak).

These considerations have led me to understand duration as time platform upon which all speech parameters operate, instead of a main phonetic parameter for distinguishing

phonemes. Duration essentially directs our attention to statistically significant variations in speech sound segments.

After revisiting the data obtained in Simões' (1987), it became evident that variations in sound segment duration, as observed in the rules for speech synthesis below, *signal* statistically significant variations in the phonological or phonetic domain, syntactic domain, and other linguistic and extra linguistic domains. By observing these variations, one concludes that in a sense, they become predictably reduced or lengthened in discourse. These variations are first of all consequences of speech events of all sorts, that is to say pause, word novelty, linguistic context, tone inflections, pragmatics, and so forth. Duration is a predictor *because* it is the consequence of something else.

Therefore, the description of the role of duration in this study is that duration per se does not produce speech events as significantly as tone/pitch inflections or intensity/loudness variations. In other words, the dynamic of *speech sounds* will reduce or extend in time depending on the demands of the *language* being materialized.

This first experience with speech synthesis by rule basically made me aware of the presence of duration in all linguistic domains. Before that study, I was not aware of how duration permeates all linguistic domains. I only thought of duration or quantity at the sound segment level.

The following paragraphs are intended to briefly illustrate how speech synthesis by rule works in support of the claim regarding the role of duration made in this study. Simões' (1987) study is essentially an attempt to adapt to Brazilian Portuguese the model for American English designed by Klatt (1976). The adaptation of Klatt's (1976) model proposes rules like the ones below, but to reach a better prediction, other rules are necessary and the addition of new rules may result in the elimination or changes of the ones here described. The ones used here are helpful to illustrate the purpose of the model, and how duration is predictable in such models.

For Brazilian Portuguese, the median value of sound segment was used as the initialization value, namely the inherent duration of each vowel. The inherent vowel duration is expected to become reduced or extended in connected speech, according to factors capable of producing sound segment variation. Rules for vowels as they were proposed in Klatt's work originated in strings of nonsense syllables spoken in a carrier phrase. For instance, the inherent phonological durations *D* for English were derived from phrase final monosyllables ending in a voiced stop, e.g., "bag" or "big." Such a syllable structure is not possible in

Brazilian Portuguese. That is why the median value of vowel duration was used in my Brazilian Portuguese adaptation of the model.

The rules in (SIMÕES, 1987, p. 114-117) were proposed for the prediction of the variations of duration *D* of the three PM's cardinal vowels [i, a, u] in connected speech.

In the linear equations below,

D_o stands for duration output sought at any given point of the speech;
K is a constant value for each phonological environment;
D_i is the inherent duration for each sound segment;
D_{min} is the minimum reduction the inherent sound segment duration can have, which is especially important in the production of stressed vowels.

Therefore, we obtain

equation (1) $D_o = K * (D — D_{min}) + D_{min}$

Equation (1) operates cyclically from domain 1 up to domain 4. The values for *K* are obtained by means of

equation (2) $K = D_o — D_{min} / D_i — D_{min}$.

The inherent duration of each vowel initializes each process in any position within a word. The subsequent D_i values are the outputs of the application of the rule just applied.

Level 1. Sound segment domain

Rule 1: INITIALIZATION. From the inventory the inherent sound segment duration is set. D_i of PM's vowels [i, a, u]:

[i] = 56 ms (7.0 mm) [a] = 64 ms (8.0 mm) [u] = 48 ms (6.0 mm)

Rule 2: If the phonetic voiceless fricative [“s] follows the vowel within the same syllable, shorten the vowel by *K* = -1.17 (65 % decrease). In case of the consonant [x], shorten the vowel by 25%, i.e. *K* = *.17*

Rule 3: If a phonetic voiced consonant follows the vowel within the same syllable, no change, *K = 1.*

Level 2. Word domain

Rule 4: If the vowel is in postonic (post1) position, decrease the vowel by 25%, *i.e. K = .17*

Rule 5: If the vowel is in pretonic (pr2) position, not preceded by a consonant, decrease it by 10 %, i.e. *K* = .67; if the vowel is preceded by a consonant, increase it by 42 %, i.e. *K* = 2.4

Rule 6: If the vowel is in immediate pretonic (pr1) position, increase it by 13%, *i.e. K* = 1.43

Rule 7: If the vowel is in stressed position, increase the vowel by 90 %, *i.e.* *K* = 4.

Level 3. Sentence domain

Rule 8: If the vowel is at the beginning of a sentence or a pause, no change, *K = 1.*

Rule 9: If the vowel is in sentence medial position, decrease by 13%, *K=.57*

Rule 10: If the vowel is in sentence final position without physical pause, increase the vowel by 20 %, *K* = 1.67; if the vowel is in a major final position, and a physical pause follows, increase the vowel by 32 %, *K* = 2.07 *(Rule 10 does not apply to vowel in post1)*

Level 4. Semantic domain

Rule 11: If vowel is within a focused word, increase the vowel by 60 %, $K = 3$.

Rule 12: If the vowel is in an exclamatory word, increase the vowel by 80 %, *i.e.* $K = 3.7$

(Rules 11 and 12 do not apply to vowel in post1)

The insightful revelation of these rules is the overall view that they offer of how spoken language may operate. Variations in the simple and fundamental parameter duration give a promising view for explanation of how all linguistics domains are connected through phonological rules. According to these phonological rules in PM's speech, variations in sound segment duration correlate to lexical or sentence stress, to word novelty in a sentence, to phrase boundaries, ad infinitum.

These are the rules for the four domains studied. Adding domains, i.e. factors, will change the number of rules creating a better sound segment production and better insights on information of descriptive nature.

Conclusion

Therefore, the measurements and analysis of PM's data resulted in the creation of phonological rules in four linguistic domains, but we can explore other linguistic domains to improve the contextualization of the speech analyzed.

In contemplating future research taking into account the preceding discussion, one may wonder about the range of this role of duration, that is to say the role of duration as a consequence of speech acts. By the same token, one may ask if such a role does or does not apply to sound segments. Then, the duration of English vowels for instance, could be a result of particular gestures inherent to the so-called *long and short vowels* of English. Then, the lengthening or shortening of English would not be distinguishing parameters but consequences of parameters such as tone inflexions or something else.

There are precedents to this possible explanation of tone inflections being the main factor in characterizing vowels. Romance language grammars tend to use the term *tonic* following a Greek and then Latin tradition, and its derivative terminologies *oxytonic*, *paroxytonic* and *proparoxytonic*, to describe patterns of stress in words.

Furthermore, if duration is a consequence of variations in sound segments, then speech by rule may not need to have rules for variations in duration. If speech events are adequately produced by rule during synthesis, then the time they take to become realized should happen automatically, as a result of related gestures, just as it happens in spontaneous speech. In other words, a speech by rule program would need to simply let time run normally while gestures take place.

This study is a preliminary step to explain why there are variations in sound segment duration. This is an old question that I have had, since it was first brought to my attention by Bertil Lyberg who, in the eighties, reminded me (personal communication) that "one thing is to obtain a model that accurately describes the data points, but it is another to find an *explanation* of why segment duration is lengthened or shortened in certain positions." Therefore this analysis essentially searches to explain the role of duration, something not adequately discussed in Simões (1987).

Bibliography

DELATTRE, P. C.; LIBERMAN, A. M.; COOPER, F. S. Acoustic loci and transitional cues for consonants. In *Journal of the acoustical society of America,* 27, 769-74, 1955.

FANT, G. *Acoustic theory of speech production,* 2nd. edition. The Hague: Mouton, 1970.

FRY, F.B. Duration and intensity as physical correlates of linguistic stress. In *Journal of the acoustical society of America,* 27, 4, 765–68, 1955.

JACKOBSON, R.; FANT, G.; HALLE, M. *Preliminaries to speech analysis*, 8th ed. 1969. Cambridge, Ma.: MIT Press, 1952.

KLATT, D. H. Vowel lengthening is syntactically determined in a connected discourse. In *Journal of phonetics,* 3, 129-40, 1975.

_____. Segmental duration in English. In *Journal of the acoustical society of America,* 59, 1208-21, 1976.

KOZHEVNIKOV, V.A.; CHISTOVICH, L.A. Speech articulation and perception. *JPRS 30*. Washington, DC, 1965.

LAZARIDIS, A.; GANCHEV, T.; MPORAS, I.; DERMATAS, E; FAKOTAKIS, N. Two-stage phone duration modelling with feature construction and feature vector extension for the needs of speech synthesis. In *Computer Speech and Language* 26, 274–92, 2012.

LEHISTE, I.; PETERSON G.E. Transitions, glides, and diphthongs. In *Journal of the acoustical society of America*, 33, 268-77, 1961.

LINDBLOM, B.; LYBERG, B.; HOLMGREN, K. *Durational patterns of Swedish phonology: do they reflect short-term motor memory processes?* Indiana, Bloomington: Indiana University Linguistic Club, 1981.

MARTINET, A. Phonology as functional phonetics. *Publications of the Philological society*, no. 15. London: Oxford University Press, 1-27, 1949.

MILLER, G. A.; NICELY, P.E. An analysis of perceptual confusion among some English consonants. In *Journal of the acoustical society of America*, 27 : 339-352, 1955.

NOTEBOOM, S.G. Temporal patterns in Dutch. In *Proceedings of the 7th international congress of phonetic sciences*, 984-89, 1972.

PIKE, K. L. *The intonation of American English*. Ann Arbor: University of Michigan Press, 1945.

SIMÕES, A.R.M. *Temporal organization of Brazilian Portuguese vowels in continuous speech: an acoustical study*, Ph.D. diss. Austin, Texas, USA: University of Texas, 1987.