Journal of Phonetics (1996) **24,** 445–473

Pitch downtrend in Spanish

Pilar Prieto and Chilin Shih

Speech Synthesis Research Department, Bell Laboratories, Lucent Technologies, 700 Mountain Avenue, Murray Hill, NJ 07974, U.S.A.

Holly Nibert

University of Illinois at Urbana-Champaign, U.S.A.

Received 20th December 1994, and in revised form 3rd July 1996

In this study, the scaling of peak fundamental frequency (f_0) values in Mexican Spanish downstepping contours is examined as a function of the following linguistic factors: (1) phrasal length; (2) temporal distance between pitch accents; (3) phrasal position; and (4) f_0 value of preceding peak. The motivation for this study stems from contradictory claims in the literature regarding whether downtrend is governed by local or global factors. Three speakers of Mexican Spanish read a total of 540 declarative utterances (2304 target pitch accents) of varying length (from two to five pitch accents) and varying distance between H* pitch accents (from two to three intervening unstressed syllables). The data reveal that the f_0 value of the previous peak (as opposed to phrasal position) is the most important predictor of peak height. In our data, between 65 and 80% of the variance of the data is predicted by exclusively using a local *downstep ratio* or constant reduction in the previous peak's pitch value. Neither phrasal length nor distance between adjacent pitch accents has a significant effect on the height of a given f_0 peak. Utterance-final peaks are best predicted by using a particular ratio of decay (higher than the downstep ratio) *and* a phrasal length factor: the use of the latter factor reflects a tendency for final peaks in longer utterances to remain at a relatively high f_0 level.

1. Introduction

This article examines two fundamental questions regarding the behavior of pitch downtrend (i.e., the general lowering of pitch over the course of a phrase) in Mexican Spanish. The first question addresses the nature of downtrend in this dialect. The term *declination* was widely used to refer to the tendency of fundamental frequency to gradually decline over the course of an utterance, and was traditionally viewed to have a physiological basis: the continuous f_0 decay observed in many languages was believed to be triggered by an automatic physiological mechanism (Fujisaki, 1983, 1988 for Japanese; Thorsen, 1980 for Danish; Bruce, 1977 for Swedish; Lieberman, 1967; Liberman, 1975; Cooper & Sorensen, 1981 for English, among many others). It was predicted that the temporal *distance* between two equally prominent adjacent peaks in a f_0 contour would

0095-4470/96/040445 + 29 $25.00/0

significantly affect the pitch level attained by the second peak: the greater the distance between the two peaks, the lower the f_0 value of the second one. Pierrehumbert's (1980) and Liberman & Pierrehumbert's (1984) work represented a major breakthrough in the study of declination: they discovered the stability of f_0 peaks in descending contours under varied prominence conditions, and successfully modeled such peaks by using a constant ratio of decay. They found that time-dependent lowering was almost absent in their data and that pitch descent could be modeled by an accent-by-accent decay (which they termed *downstep*), eliminating the need for a global declining component. Thus, they proposed that the continuous downtrend observed in certain contours in English (i.e., *downstepping* contours) could be explained as the deliberate use of the step accents by the speaker. After Liberman & Pierrehumbert's (1984) contribution, little evidence has been given to support the existence of a time-dependent *declination* effect. Early studies attempting to describe a global declination effect generally compared target f_0 points that were not equivalent phonologically, making it hard to tease out the real effect of time on the descent of the pitch accents. Still, recent models such as Pierrehumbert (1980) and Fujisaki (1983, 1988) continue to assume that both mechanisms are active, namely, a local downstep mechanism (which applies accent-by-accent) and a global declination line that is fitted to the entire phrase. One of the goals of this experiment is to investigate the influence of the temporal distance between downstepped peaks on f_0 peak height in Mexican Spanish to determine whether it is controlled by a gradual time-dependent *declination* effect, a ratio-driven *downstep,* or both. If time-dependent declination is active, we would expect a consistent amount of lowering as the temporal distance between accents increases.

The second question concerns the effect of phrasal length on the height of successive peaks in the f_0 contour. One of the controversial issues in intonational studies is whether the speaker plans f_0 contours at the phrase level or at a more local level (*hard vs. soft preplanning*, in Liberman & Pierrehumbert, 1984). Recent instrumental studies make contradictory claims about the relationship between the length of an utterance and the height of the first peak (for a review, see Ladd & Johnson, 1987): while authors like Cooper & Sorensen (1981) reported a significant increase in peak height in longer utterances, other authors found that peak values are more or less constant in a given position, regardless of utterance length (Thorsen, 1981: 42; Sternberg, Wright, Knoll & Monsell, 1980; Liberman & Pierrehumbert, 1984; van den Berg, Gussenhoven & Rietveld, 1992).[1] The second goal of the present study is to examine the influence of phrasal length (measured in terms of number of pitch accents per utterance) on the scaling of f_0 peaks.

Even though it has been observed that f_0 peaks in Spanish declarative sentences tend to descend throughout the sentence (Fant, 1984), no attempt has been made to identify the factors affecting the scaling of peaks in this language. The present

[1] Ladd & Johnson (1987: 240) attempted to explain such contradictory results by suggesting that branching structure could be involved in boosting the pitch level of the first peak. They pointed out that "the apparent effect of utterance length on the first peak (P1) is in reality largely an effect of whether the firt major constituent of the utterance contains more than one accented item". In particular, "if utterance length is increased by adding accents to the first major constituent (..), then there will be a substantial effect on P1". For example, compare the following two sentences: "Graham Alexander was promoted to sergeant" *vs.* "Graham was rapidly promoted to sergeant". In the first sentence, P1 is claimed to be notably higher, since the first constituent contains more than one pitch accent.

study provides a preliminary descriptive model of peak height in this language, and analyzes the relationship between peak height, phrasal position and other prosodic conditions. Also, a phonetically-explicit model of Spanish descending contours can lead to improvement of existing f_0 assignment algorithms for text-to-speech purposes.

The analysis of a total of 2304 target pitch accents included in read declarative utterances of varying length (from two to five pitch accents) and varying distance between pitch accents (from two to three intervening unstressed syllables) indicates the following. First, starting and ending f_0 values are more or less constant over utterances of varying phrasal length. Second, peaks in a given position in the utterance show a near constant f_0 value, regardless of the number of syllables separating them or the length of the utterance. In general, our data provide evidence that the effect of time distance between peaks in the speech of our three informants is very small and statistically not significant. Peak height is successfully predicted by a constant speaker-dependent *downstep ratio*, with no need to resort to a global declining line. Similarly, our data show a lack of utterance preplanning of f_0 contours on the part of our three speakers. Finally, utterance-final accents are affected by a strong final lowering effect.

2. Experimental design

The scaling of f_0 peaks in read declarative utterances was studied as a function of the following factors: (1) utterance length (in terms of number of pitch accents per utterance); (2) temporal distance to the previous accent (in milliseconds or in number of unstressed syllables); (3) utterance position; and (4) value of the previous peak. These four factors were initially chosen because they have been shown to have an overall effect on f_0 scaling in different languages.

Figs. 1 and 2 illustrate the f_0 contour under study, a typical downstepped f_0 contour of an utterance with five pitch accents, as produced by speakers RS and JC, respectively. These contours could be described as series of simple peaks, such that each peak is lower than the one preceding it. To characterize the pitch contour phonologically, we adopt Pierrehumbert's (1980) theory of intonational structure. Within this framework, the pitch accents in question could be characterized either as a series of H* + L accents or as a series of downstepped H* accents. The first option

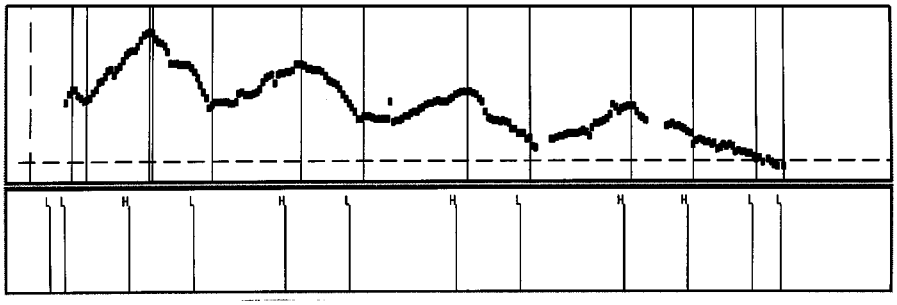

Figure 1. f_0 contour of the five-accent utterance *Rayo de luna de mayo de gala de Lola* as realized by speaker RS.

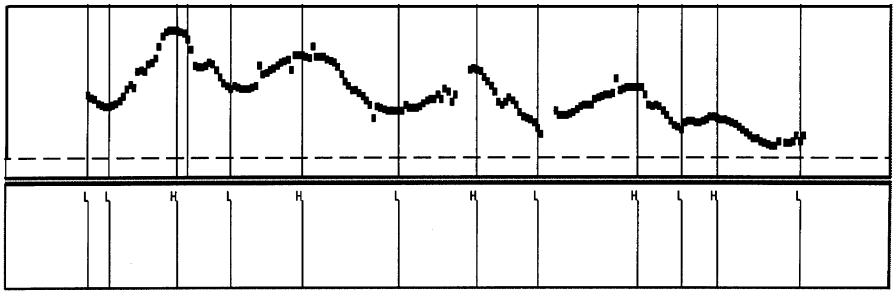

Figure 2. f_0 contour of the five-accent utterance *Rayo de luna de mayo de gala de Lola* as realized by speaker JC.

(chosen by Pierrehumbert (1980) for English) indicates that the trigger of accent stepping is the presence of the L valley after the rise. This option does not seem to be adequate in our case, however, since different varieties of Spanish have been reported to have a phonological distinction between a H* + L and a H* accent: in Caracas Spanish, for example, while H* is employed in neutral declaratives, H* + L is used to convey a complaint or resentment (Sosa, 1991: 15, 110). Given the lack of an exhaustive study of pitch accents in Mexican Spanish and the possibility that H* + L could behave as a phonologically independent accent, we will not use H* + L to describe downstepping accents. Instead, we assume that downstep in this variety of Spanish is simply the result of the phonological choice of the speaker, and that the pitch accent involved is !H* (Prieto, van Santen & Hirschberg, 1995), a downstepped H* pitch accent within the ToBI labeling scheme (Pitrelli, Beckman & Hirschberg, 1994). In the labeling scheme used for our data analysis and figures (discussed in detail in Section 2.2), we make use of the symbol L, however not as a pitch accent. We use it for lack of a better way to refer to the lowest absolute f_0 value between pitch accents. We are not claiming that a phonological L* underlies these points; L is simply a measurement tick mark allowing us to monitor these values. Finally, it is not the goal of this paper to model the L values in these contours.[2]

2.1. *Test utterances*

The target database consisted of 30 phrases derived from all possible combinations of two to five pitch accents separated by two to three intervening unstressed syllables. For example, the following utterances (written orthographically, marking the lexical stress of each content word) illustrate how the number of intervening unstressed syllables was systematically increased for two and three-accent phrases. As shown below, the utterances consist of a noun phrase modified by an increasing number of embedded prepositional phrases; this type of syntactic structure was used because it naturally elicits the type of contour under study and it allows for easy control of the segmental and syntactic structure of the utterances. To facilitate comparisons of pitch accents in the same position across utterances of different

[2] We are currently undertaking a phonetic analysis of the L values present in the same contours (Prieto, in prep).

length, the ordering of content words in each phrase is kept constant and the placement of the accent is always in the first syllable of a bisyllabic noun. Finally, to minimize the segmental effects on f_0, all content words contain only sonorant consonants (with the exception of the word *gala*).

Appendix 1 includes a complete list of the phrases comprising the database.

TWO ACCENTS

1. *ráyo de lúna* "ray of moon"
2. *ráyo de la lúna* "ray of the moon"

THREE ACCENTS

1. *ráyo de lúna de máyo* "ray of moon of May"
2. *ráyo de lúna de mi máyo* "ray of moon of my May"
3. *ráyo de la lúna de máyo* "ray of the moon of May"
4. *ráyo de la lúna de mi máyo* "ray of the moon of my May"

2.2. *Recording and measurements*

Three speakers of Mexican Spanish[3] produced the target declarative utterances in a single intonational phrase avoiding emphasized constituents. The utterances were randomized and read six times at a normal speech rate, for a total of 540 utterances. Utterances containing discontinuities or unsolicited prominence effects were discarded. For each utterance, the f_0 contour was extracted using the Waves speech analysis package (Talkin, 1989), and relevant f_0 measures were taken (see Fig. 1), avoiding f_0 tracking errors and local segmental effects. The labeling scheme included the following key points of the f_0 contour:

1. Phrase-initial f_0 value (marked with L)
2. Highest absolute f_0 value for each pitch accent (H)
3. Lowest absolute f_0 value between pitch accents (L)
4. Phrase-final f_0 value (L)
5. Beginning of each accented vowel, to be used as a time reference

As shown in Figs. 1 and 2, our speakers uttered the final part of a declarative contour in different ways: while speakers JC and AH produced very compressed H* accents in utterance-final position (see Fig. 2), speaker RS (Fig. 1) realized the final accent as a continuously falling slope, with no clear pitch excursion. The consistently different shape of RS's final accent caused us to redefine the meaning of the labels in the prelabeling state for RS, since they are rendered senseless in this context: it makes no sense to employ the H label to refer to the highest point of the f_0 peak, since in this context no peak is present. In this case, the last H corresponds to the onset of the accepted syllable and the pennultimate L corresponds to the end of the accepted vowel. Thus, for RS's final accent, the labels H and L signal important segmental boundaries in the phrase-final accented syllable rather than values of dynamic points in the f_0 contour. The bottom part of Fig. 1 illustrates the labeling scheme used for RS with a five-accent phrase. The f_0 values of the labeled points were plotted to produce graphs, and any remaining f_0 tracking errors revealed in the graphs were corrected.

[3] JC, AH, and RS are from different regions of Mexico. JC and AH come from Mexico City, and RS from Ciudad Juárez, located in the northern state of Chihuahua.

We contend that the different shapes observed in the last pitch accent represent optional phonetic manifestations of the same final accent, and do *not* reflect a prominence increase or a distinct linguistic use. Phonetic variation in the production of the last pitch accent of Spanish declarative sentences has also been observed in Peninsular Spanish (Fant, 1984: 36) and in different varieties of Latin American Spanish (Delattre, Olsen & Poenack, 1962; Navarro-Tomas, 1994; Sosa, 1991), where sentence-final pitch accents are produced either as very compressed peaks or with no peak at all. In fact, Delattre *et al.* (1962: 237) point out that ''both the form and the manner of the pitch contour for finality in American English and Mexican Spanish are different. On the stressed syllable, American rises, then descends slowly; Spanish, without having risen, descends abruptly''.

We conducted a small perceptual experiment to examine the degree of perceived prominence of pitch accents in the recorded utterances and test whether listeners are able to identify different levels of prominence of the final accents across the three speakers. The stimuli consisted of 12 utterances of different lengths from each speaker (RS, JC, and AH), for a total of 36 utterances (see Appendix 2). Four native speakers of Spanish heard the randomized utterances from the three speakers in three separate sets and classified the relative prominence of each accent on a scale of 1 to 5. Table I shows the mean prominence score of final *vs.* non-final accents as rated by our four listeners, together with the ratio of the final accent and all non-final accents, given in parentheses. The scores of all non-final accents are combined because there is no consistent pattern.

Consistently, utterance-final accents were classified to have a lower degree of perceived prominence than non-final accents. Yet the fact that relative prominence scores for final accents show no substantial differences across speakers indicates that RS's final accent with no pitch excursion represents a case of a clear accent perceptually equivalent to the final peak of the other two speakers with compressed peaks. The mean relative prominence ratios of final peaks and non-final peaks are

TABLE I. Mean prominence score (on a 1 to 5 scale) of accents in final *vs.* non-final positions for our three speakers (RS, JC, and AH). The relative prominence of the final accent (with respect to all preceding accents) is given in parentheses

	Prominence scores (1 to 5)		
Position	Speaker RS	Speaker JC	Speaker AH
Listener PG			
Non-final	3.12	3.53	3.37
Final	2.5 (0.80)	2.66 (0.75)	2.83 (0.83)
Listener EL			
Non-final	4.62	4.53	4.40
Final	4.08 (0.88)	4.16 (0.91)	4.08 (0.92)
Listener JP			
Non-final	3.18	4	4.03
Final	2.19 (0.91)	4 (1)	3.33 (0.82)
Listener JS			
Non-final	3.46	3.62	3.5
Final	2.58 (0.74)	3.25 (0.89)	3 (0.85)

0.83 for speaker RS, 0.88 for speaker JC, and 0.85 for speaker AH. It is expected that listeners normalize peak height by lowering effects (Pierrehumbert, 1979), so that a pitch peak that is lower in absolute f_0 than the previous peak can be perceived as being equally high as the previous one. Our study goes one step further and shows that listeners are capable of compensating for the total absence of pitch excursion in making a judgment on the prominence level of the utterance-final pitch accent, presumably because this particular mode of pitch realization is sanctioned by the intonational phonology of Spanish. Section 3.4. deals with the modeling of utterance-final accents in our data.

3. Results

3.1. *Utterance-initial and utterance-final f_0 values*

In our data, starting and ending f_0 values are nearly constant for a given speaker. Fig. 3 shows the mean values of utterance-initial (left panel) and utterance-final (right panel) f_0 values for phrases of different lengths (from two to five pitch accents). Utterance-initial f_0 values are invariant for all three speakers when phrasal length varies from two to five accents. One-way ANOVAs for each speaker show no significant effect of phrasal length (where $p \leqq 0.01$ is the level of statistical significance): RS: $F(3, 164) = 2.55$, $p = 0.057$, JC: $F(3, 163) = 1.36$, $p = 0.26$, and AH: $F(3, 176) = 3.09$, $p = 0.029$.

Utterance-final f_0 values are not affected by phrasal length for two of the three speakers: JC: $F(3, 162) = 2.48$, $p = 0.063$, and AH: $F(3, 176) = 0.76$, $p = 0.52$. For RS ($F(3, 164) = 4.59$, $p < 0.01$), final f_0 values are slightly higher when sentences are shorter. However, mean f_0 values from different phrasal length conditions are very close, ranging from 74 Hz to 76 Hz.

These results are similar to those found by Liberman & Pierrehumbert (1984: 181) for English: in general, utterance-initial and utterance-final f_0 values were nearly constant for a given speaker and were not correlated with utterance length or with the values of f_0 peaks (pronounced with different degrees of emphasis). On the other hand, Thorsen's (1981) results for Danish showed a tendency to have lower ending points in longer utterances. Yet, she points out that in utterances with more than

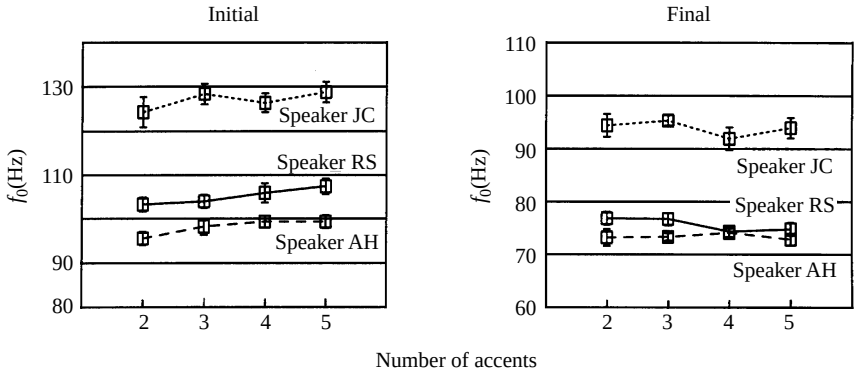

Figure 3. Mean values (in Hz) of utterance-initial (left) and utterance-final (right) f_0 values of phrases of different lengths (two to five pitch accents). The height of the bars represents standard errors.

TABLE II. Mean values (in Hz) of the pitch range of the first peak in utterances of different lengths (three to five pitch accents)

	Speaker RS		Speaker JC		Speaker AH	
Accents	f_0 range	SD	f_0 range	SD	f_0 range	SD
3	30.06	2.09	49.16	2.61	6.06	1.47
4	28.81	2.41	48.42	2.31	2.55	1.53
5	28.79	2.57	53.46	2.48	3.41	1.61

four stress groups, end points tend to remain constant, ''which is probably a reflection of a physiological constraint''.

3.2. *Pitch range*

In order for pitch accents in our data to be comparable, they need to have been realized with the same degree of prominence, since pitch range variation represents an additional factor that could confound the effects of the factors under study. As the initial f_0 values are more or less constant for a given speaker, the initial f_0 rise (measured as the difference in Hz from the lowest f_0 point at the beginning of a contour to the highest f_0 value of the first peak) should be a good measure of the degree of prominence with which the first pitch accent was produced. Table II shows the mean range of the first peak and standard error values in utterances of three to five accents for the three speakers. The rather constant pitch range values of the initial rise, together with their low standard error values, indicate that our utterances were produced in a more or less constant pitch range and that H points are therefore comparable.

3.3. *High f_0 peaks*

This section examines the effect of the following factors on the scaling of f_0 peaks (1) temporal distance between pitch accents (in both number of unstressed syllables, varied from two to three, and in time), and (2) phrasal length in terms of number of pitch accents (varied from two to five). As mentioned before, examination of the first factor reveal whether or not a time-dependent declination effect can be found in these Spanish contours, while the latter factor will be a measure of whether or not there is evidence for preplanning of f_0 contours at the phrase level.

3.3.1. *Effect of temporal distance between peaks*

This section examines how time affects the f_0 scaling of H peaks, that is, whether or not there is evidence for a global time-dependent descending f_0 component or *declination.* The basic question motivating the study in this section is whether there exists a declination effect in Spanish independent of the downstep effect, to be discussed in Sections 3.3.3. and 3.3.4. The term declination has been used to refer to two very different concepts. One definition includes all of the downtrend observed in a sentence, as in Cooper & Sorensen (1981), Lieberman (1967), and Fujisaki (1983). The other restricts the term to a global, time-dependent downtrend after other

lowering effects as downstep or final lowering have been factored out, as in Pierrehumbert (1980), Liberman & Pierrehumbert (1984), Poser (1984), and Prieto, Nibert & Shih (1996). As mentioned before, this paper follows the second definition, namely, declination is considered to be the residue downtrend after all other accountable lowering effects have been factored out.

To test the declination hypothesis, we included the following two conditions in the experimental design: each pair of peaks is separated by either two or three unstressed syllables (NUMSYLS = 2 and NUMSYLS = 3). We will be comparing only phonologically equivalent events, namely, H* pitch accents having undergone the phonological rule of downstep such that each accent's f_0 peak is lower than the one preceding it. In this contour, successive f_0 peaks decrease in height due to downstep. If declination is present in addition to downstep, f_0 reduction is expected to be greater as the time interval between peaks increases, that is, second peaks in the NUMSYLS = 3 condition should be lower than those of the NUMSYLS = 2 condition (since there is a larger time interval between the two peaks of NUMSYLS = 3). It is also expected that the amount of pitch drop correlates with the time distance between both peaks. Otherwise, if there is no declination effect in our data, the second peaks in the two conditions should be subjected to the same amount of lowering from downstep, and therefore show no difference in peak height and/or the amount of pitch drop.

Since final peaks have different realizations for the three speakers (the final peak is not measurable for speaker RS) and have different lowering ratios than non-final peaks (see Section 3.4.1.), we excluded the final peak from the study of declination presented in this section, even though results obtained from data with final peaks are very similar to the ones presented below with final peaks excluded.

The three plots in Fig. 4 give an overview of the relationship between peaks separated by two and three unstressed syllables: we plot the mean absolute f_0 values (in Hz) of four successive f_0 peaks (H1 to H4) for our three speakers. Peaks separated by two unstressed syllables (NUMSYLS = 2) are plotted with solid lines and peaks separated by three unstressed syllables (NUMSYLS = 3) are plotted with dotted lines. The three plots in Fig. 4 show that in general, peaks in the same phrasal position are strikingly similar, and that the difference between the two conditions is quite small. Also, not all patterns are consistent with the prediction of declination, where the dotted lines are expected to lie below the solid line. It appears that the data from speaker JC are the only data that conform with the declination hypothesis.

For easy comparison of the peaks in the conditions NUMSYLS = 2 and NUMSYLS = 3, Fig. 4 does not include the time dimension. Results in Table III demonstrate that, for the three speakers, changing the syllable count has a substantial effect on the inter-accent separation within groups: mean durational distance is significantly different between NUMSYLS = 2 and NUMSYLS = 3 for all speakers, showing that three intervening unstressed syllables indeed take longer time to utter than two intervening unstressed syllables, even though the duration of each syllable is shorter in the NUMSYLS = 3 condition. The time difference between the two conditions ranges from 97 ms for JC and AH, to 163 ms for RS, which we believe is substantial enough to support a declination test on our data.

Table IV shows the mean peak height of accents preceded by two (NUMSYLS = 2) and three unstressed syllables (NUMSYLS = 3). In general, no significant differences

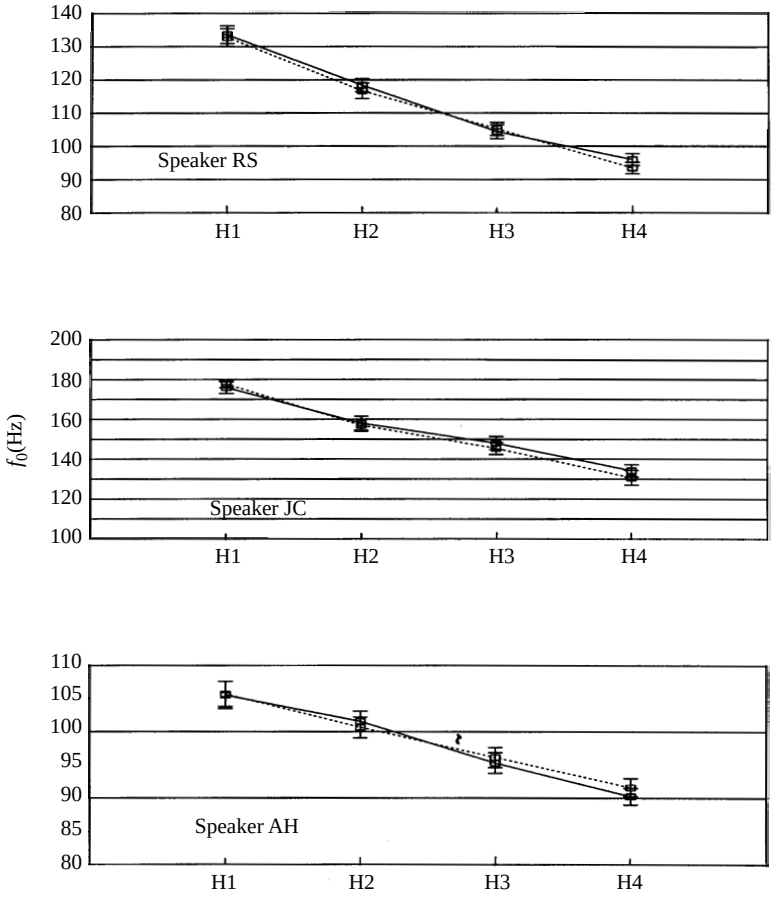

Figure 4. Mean f_0 values of adjacent peaks (H1 to H4) corresponding to an increase in the number of unstressed syllables between pitch accents (two to three) for the three speakers. Solid lines link the peaks separated by two unstressed syllables (—, NUMSYLS = 2) and dotted lines link those separated by three unstressed syllables (· · · ·, NUMSYLS = 3). The height of the bars represents standard error values.

TABLE III. Mean durational distance (in ms) and standard deviations between peaks with two intervening unstressed syllables (NUMSYLS = 2) and with three intervening unstressed syllables (NUMSYLS = 3)

	NUMSYLS = 2	SD	NUMSYLS = 3	SD	*t*-test (df)
RS	507.19	63.53	670.01	65.54	$p < 0.001$ (370)
JC	432.46	57.23	529.30	55.84	$p < 0.001$ (371)
AH	483.69	44.59	580.49	47.37	$p < 0.001$ (406)

are found between peaks in the NUMSYLS = 2 and NUMSYLS = 3 populations, for all three speakers. The third speaker (AH) shows a trend that contradicts the predictions of the declination hypotheses, namely, the second peak is higher when the time interval from the previous peak is longer.

TABLE IV. Mean and standard deviations of f_0 values of the second peak preceded by two (NUMSYLS = 2) and by three (NUMSYLS = 3) unstressed syllables

	NUMSYLS = 2	SD	NUMSYLS = 3	SD	*t*-test (df)
RS	111.81	9.10	110.14	9.23	$p = 0.08$ (370)
JC	153.68	11.56	152.12	11.68	$p = 0.196$ (371)
AH	98.85	5.62	99.28	5.16	$p = 0.423$ (406)

TABLE V. Speaker RS: Mean and standard deviations of f_0 difference between peaks with two and with three intervening unstressed syllables

	POS2	SD	POS3	SD	POS4	SD
NUMSYLS = 2	14.56	6.78	11.72	2.74	6.54	3.62
NUMSYLS = 3	15.69	4.86	12.04	4.97	8.66	3.08
t-test (df)	$p = 0.093$ (153)		$p = 0.52$ (130)		$p < 0.01$ (83)	

Since the distance in time between the two test conditions is small, the natural fluctuation of f_0 values in speech could be hiding a possible effect of time on f_0 downtrend. It is also clear from the trajectory plots in Fig. 4 that phrasal position has an effect on peak height: we find more pitch drop at the beginning of the utterance (see discussion on downstep in later sections) than later in the utterance. Thus, it is probably more reasonable to compare the amount of pitch drop from one accent to the next in the NUMSYLS = 2 condition and the NUMSYLS = 3 condition, in different positions in the utterance. Tables V, VI, and VII show the mean distances in f_0 between pairs of adjacent peaks separated by two or three unstressed syllables in different positions in the utterance (H1–H2 = POS2; H2–H3 = POS3; and, H3–H4 = POS4). Standard deviations and t-test results show that even though there is a tendency for NUMSYLS = 3 to have more pitch drop than NUMSYLS = 2, the results are not statistically significant. Nearly all pairs of NUMSYLS = 2 and NUMSYLS = 3 are

TABLE VI. Speaker JC: Mean and standard deviations of f_0 difference between peaks with two and with three intervening unstressed syllables

	POS2	SD	POS3	SD	POS4	SD
NUMSYLS = 2	16.57	8.63	7.67	5.18	9.57	4.56
NUMSYLS = 3	19.18	7.11	10.00	5.67	11.77	5.15
t-test	$p = 0.044$ (153)		$p = 0.016$ (129)		$p = 0.038$ (85)	

TABLE VII. Speaker AH: Mean and standard deviations of f_0 difference between peaks with two and with three intervening unstressed syllables

	POS2	SD	POS3	SD	POS4	SD
NUMSYLS = 2	3.30	3.76	4.97	3.17	4.19	2.62
NUMSYLS = 3	4.20	3.84	4.32	3.49	3.81	4.20
t-test	$p = 0.13$ (166)		$p = 0.25$ (142)		$p = 0.60$ (94)	

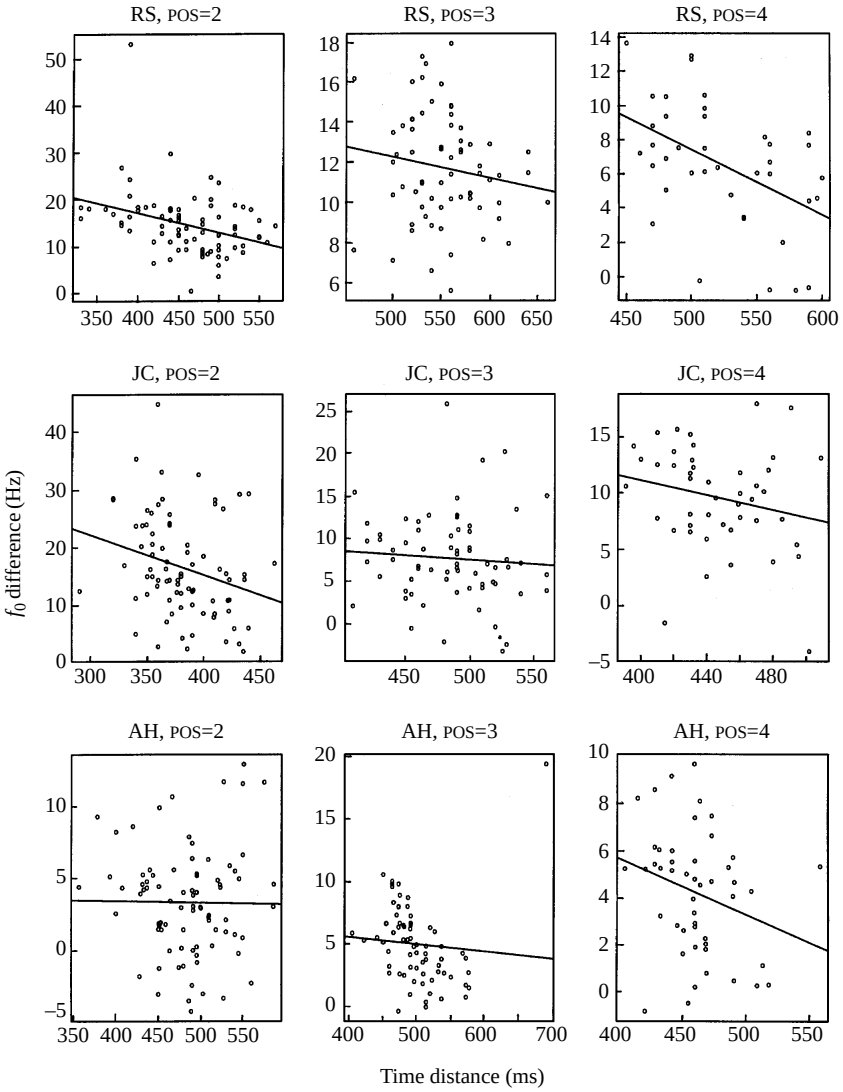

Figure 5. Mean f_0 difference between peaks in the NUMSYLS = 2 condition in different positions in the utterance (POS2, POS3, POS4), as a function of the distance in time (ms) between the two peaks.

non-significant at the $p = 0.01$ level, except for the pair in position 4 (POS4) of speaker RS. Moreover, two of three pairs of means from speaker AH show a trend that contradicts the declination hypothesis.

There is further evidence that the general trend shown in the means of NUMSYLS = 2 and NUMSYLS = 3 *cannot* be interpreted as a product of declination: as illustrated in Fig. 5 for the NUMSYLS = 2 condition, the 3 speakers show a *negative* correlation between the distance in time separating any pair of peaks and the corresponding drop in Hz. This picture holds no matter how the data cells are split or combined. Each graph in Fig. 5 plots the amount of pitch drop (in Hz) between pairs of peaks as a function of the time distance (in ms) between those peaks, for

each position in the utterance. Pitch and time differences between first and second peaks (POS2) are plotted in the left column, differences between second and third peaks (POS3) are plotted in the middle column, and differences between third and fourth peaks (POS4) are plotted in the right column. The solid line in each plot represents the least square regression value: consistently, all lines show a negative correlation between the distance in Hz between two given peaks and the distance in time separating them. The most straightforward interpretation of declination as a global, time-dependent effect predicts that a drop in pitch between any two peaks will be *positively* (not *negatively*) correlated with the distance in time between them. Therefore, our results constitute hard evidence against this view of declination.

In sum, the results presented in this section provide very little evidence for declination. First, there is no significant difference between the two controlled populations under investigation. Second, and more importantly, our data do not show a positive correlation, as predicted by declination, between the amount of pitch drop and the time interval between any two given peaks. To the contrary, the data show a negative correlation in this respect.

3.3.2. *Effects of phrasal length*

As mentioned, some researchers have found that the height of utterance-initial peaks is significantly greater in longer utterances; it is argued that this result reflects speaker preplanning of the f_0 contour at the phrase level (Thorsen, 1980 for Danish; Cooper & Sorensen, 1981; Sternberg *et al.,* 1980 for English). In contrast, other instrumental studies report results not consistent with the notion of preplanning (Liberman & Pierrehumbert, 1984: 220).

The three plots in Fig. 6 show schematized mean f_0 contours of utterances of different lengths (from two to five pitch accents), holding the number of unstressed syllables between accents constant at three. If we compare the height of peaks in a given position in the utterance (e.g., H2) over utterances of different lengths, we find that the peaks reach a more or less constant height. The only exception is the first peak of a two-accent utterance, which is substantially lower than the first peak of longer utterances, and the utterance-final peak value.

The three plots in Fig. 7 show the mean peak height (in Hz) of peaks in different positions in the utterance (H1–H5), in phrases of different lengths (from two to five pitch accents). The data from the three speakers show that, with the exception of initial peaks in two accent utterances (see solid lines in the plot) and utterance-final peaks, mean f_0 height of a peak in a given phrasal position is similar across the various phrasal lengths. Another very interesting phenomenon displayed by the data in the three plots in Fig. 7 is the tendency for penultimate peaks to be lower than non-penultimate peaks in the same position in the utterance. In the plots, the height of the third peak of the four-accent utterance is lower than the third peak of the five-accent utterance, for the three speakers. Also, three out of the six possible comparisons are statistically significant.

Thus, in general, our data do not exhibit phrasal-length effects on the height of f_0 peaks. The height of a peak is determined by its position in the utterance, that is, the levels attained by a peak in a given position in the utterance are fairly constant, regardless of phrasal length or temporal distance between accents. Moreover, the fact that initial peaks are fairly constant across utterances of different lengths

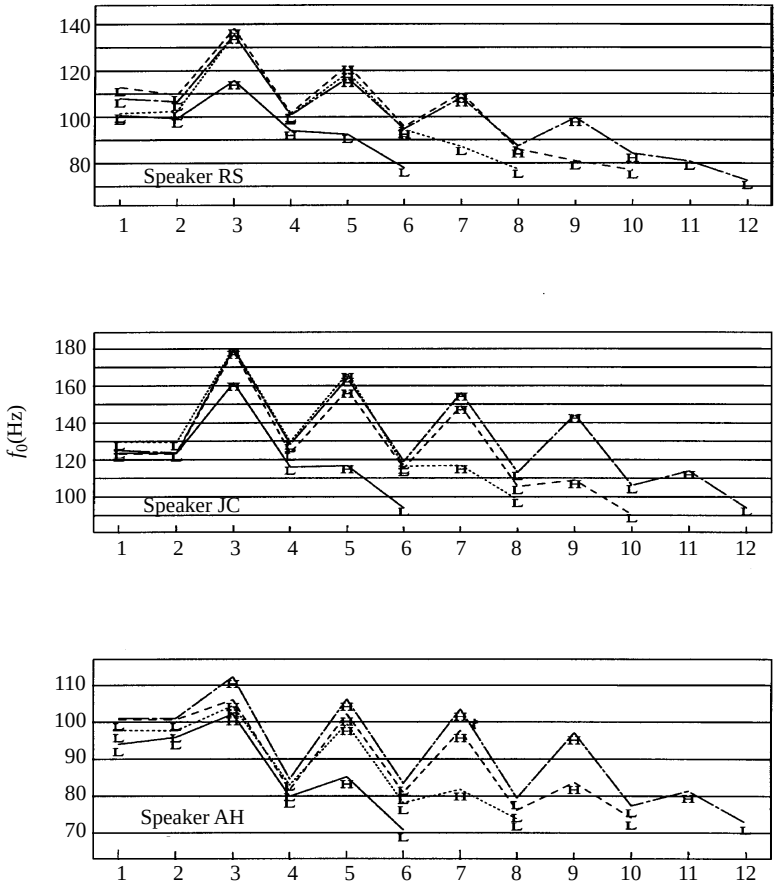

Figure 6. Mean schematized f_0 contours of five utterances of different lengths (from two to five pitch accents), keeping constant (at three) the number of intervening unstressed syllables between pitch accents, for the three speakers. —, two accents; · · · · three accents; - - - four accents; — · — five accents.

represents a preliminary indication of the lack of phrasal look-ahead in our data. The data above show that, apart from position in the utterance, an important factor in determining pitch height is the distance to the end of the utterance: in general, utterance-final and utterance-penultimate accents are affected by a lowering phenomenon, which triggers an extra amount of lowering for accents close to the end of utterances. The penultimate lowering accounts for the apparent lowering of the initial peak in two-accent utterances: we maintain that the lower initial peak is a result of final lowering, not of preplanning.

3.3.3. *Linear model of non-final peak height: Model 1*

To sort out the contribution of factors such as phrasal length and temporal distance between accents, we performed a multiple regression analysis. We initially attempted to model peak height using the following parameters: phrasal length (NUMACC: 2, 3, 4, 5 pitch accents), position in the utterance (POSSENT: 1, 2, 3, 4, 5), and distance to the previous accent (NUMSYLS: 2, 3). Due to their divergent behavior,

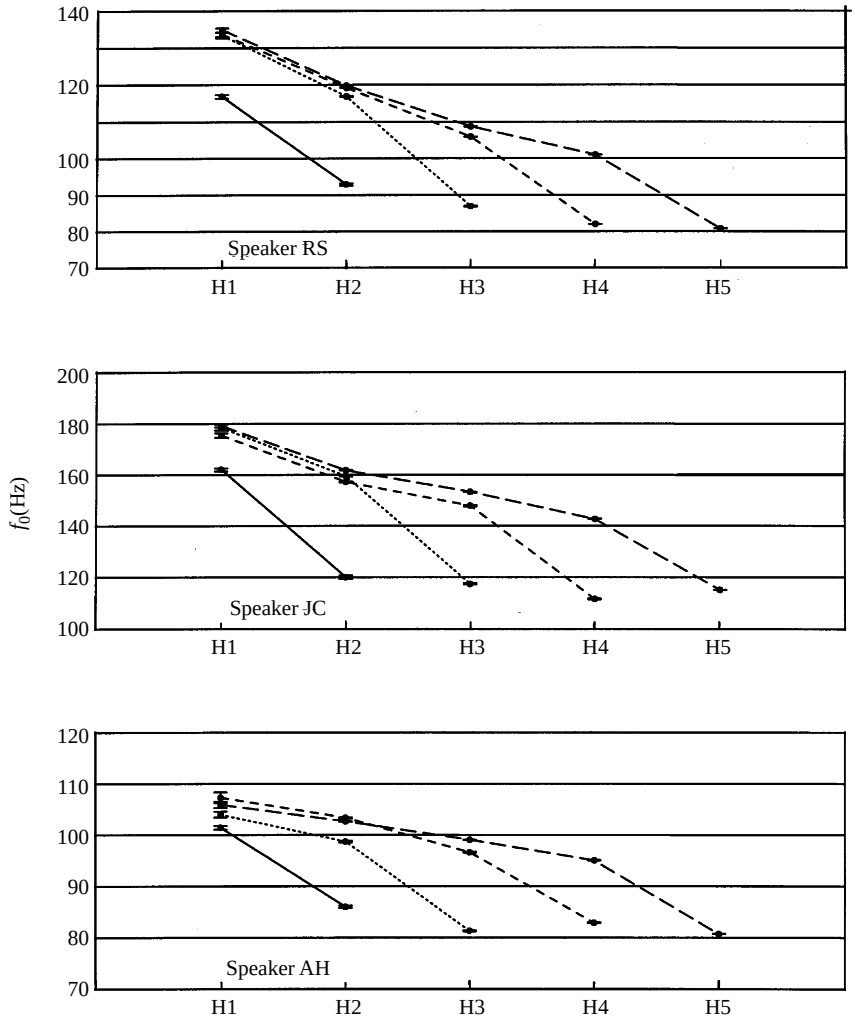

Figure 7. Mean f_0 peak values (in Hz) in different positions (H1 to H5) in utterances of different lengths (from two to five pitch accents), for the three speakers. —— Two accents; · · · · three accents; - - - four accents; –– five accents.

final peaks were initially excluded from the analysis: including utterance-final peaks in the data matrix triggers a substantial decrease in the performance of the models. Using the three factors above, the model that best fitted the data for our three speakers is shown below:

$$\text{Peak Height (JC)} = 170.97 + 2.77\ \text{NUMACC} - 9.54\ \text{POSSENT} - 1.60\ \text{NUMSYLS},$$

$$\text{Peak Height (AH)} = 102.17 + 1.64\ \text{NUMACC} - 4.12\ \text{POSSENT} + 0.352\ \text{NUMSYLS},$$

where NUMACC = Number of Accents in the Utterance (2 to 5); POSSENT = Position of the Accent in the Utterance (2 to 5); NUMSYLS = Number of Unstressed Syllables Preceding the Accent.

The predictive power of the above model is rather low: the model accounted for 64.7% of the variance for speakers RS, 37% for speaker JC, and 31.4% for speaker AH. Eliminating NUMSYLS and NUMACC decreases very slightly the performance of the model: 64.4% for speaker RS, 36.6% for speaker JC, and 31.3% for speaker AH by eliminating NUMSYLS, and 63.6% for speaker RS, 35.3% for speaker JC, 28.6% for speaker AH by eliminating NUMACC.

Thus, for the three speakers, the proportion of variance explained did not change greatly when factors such as NUMACC and NUMSYLS were removed; yet removing POSSENT caused a dramatic decrease in the performance of the model: 43% for RS, 11% for JC, and 7% for AH. This result confirms the observation that peak height in a downstepping contour is largely controlled by position in the utterance and that non-local factors like distance between accents and phrasal length have weak and inconsistent effects. Still, taking POSSENT as a main predictor, the goodness of the fit is relatively low (between 31 and 64%).

Liberman & Pierrehumbert (1984: 190) showed that the height of a given f_0 peak in English downstepping contours could be accurately predicted relative to the value of the peak preceding it by reducing the second peak by a fixed proportion. Does the height of a given peak depend on the previous peak's height, or is it targeted independently? In other words, do speakers target the height of the peaks independently of the height of preceding peaks, or is the actual target planned at a more *local* level? The data show a high correlation between the heights of two adjacent peaks within groups (peaks in a given position in the utterance, in phrases of different lengths): in general, the higher the first of two peaks, the higher the second one will be. The six plots in Fig. 8 plot the height of a given peak (H2 and H3, top and bottom panels, respectively) as a function of the height of the previous

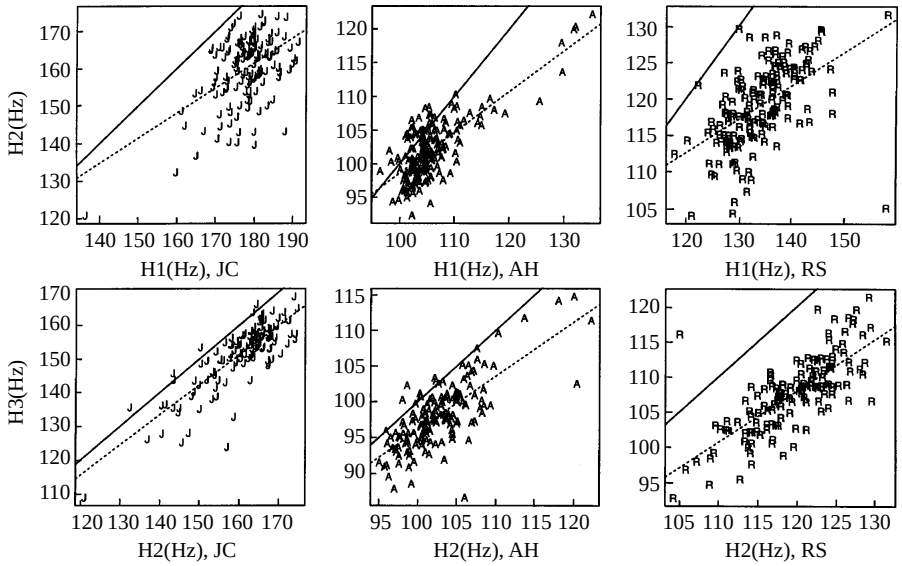

Figure 8. Peak height (in Hz) of a given peak (H2 and H3) as a function of the height of the previous peak (H1 and H2, respectively), for three speakers (R = RS, J = JC, and A = AH). The solid line plots the x = y line, whereas the dotted line is the least squared regression line.

peak (H1 and H2, respectively). The regression lines (dotted lines in the plots) summarized the positive correlation between the two variables, for each speaker. The plotting symbols R, J, and A represent the values for the three speakers (RS, JC, and AH, respectively).

Indeed, replacing POSSENT by the value of the previous peak (PREVPEAK) leads to an objectively better fit of the data: 78.7% coverage for speaker RS, 65.8% for speaker JC, and 64.6% for speaker AH. Below are the resulting regression coefficients obtained for this model, which we will call Model 1:

$$\text{Peak Height(RS)} = 25.739 + 0.692 \times \text{PREVPEAK},$$

$$\text{Peak Height(JC)} = 34.071 + 0.716 \times \text{PREVPEAK},$$

$$\text{Peak Height(AH)} = 23.93 + 0.72 \times \text{PREVPEAK},$$

where PREVPEAK = Previous Peak height (in Hz).

Thus, high correlations between pairs of adjacent peaks and the better fit of Model 1 (i.e., the one using PREVPEAK as the only predictor) clearly indicate peaks are dependent on the value of the preceding peaks, as was argued by Liberman & Pierrehumbert (1984). Thus, the speaker seems *not* to target peaks independently of previous peaks, but to control f_0 gestures at a relatively local level: if for a given reason a peak in a given position is raised, then the one following it will also be raised.

3.3.4. *Using a reference line: Model 2*

As mentioned before, Liberman & Pierrehumbert (1984) showed that downstep in English could be successfully modeled as an exponential decay to a constant nonzero asymptote (what they call the *reference* line of the speaker).[4] Phonetically, the reference line is an abstract one (in the sense that it is not realized in the pitch contour) lying somewhere in between the final f_0 value and the lowest peak. The reason for choosing a reference line in their model is to enforce a constraint on how low the peaks can go in longer utterances and to protect the peaks from falling below the f_0 minimum of the speaker. For each speaker, the reference line is scaled in such a way that it produces an optimal fit to the data by keeping a constant downstep ratio throughout all of the peaks.

The *downstep constants* (or ratio of decay between adjacent peaks) are calculated using the following equation: the ratio is the proportion of the second peak with respect to the first (scaling them with the *optimal* reference line of the speaker):[5]

$$\text{Downstep Ratio (r)} = (P(x+1) - R)/(P(x) - r),$$

where P(x) = peak height of a peak in position x; R = reference line value (optimized for each speaker).

For each speaker, we calculated an optimal reference scaling line by automatically producing the downstep ratios between mean f_0 peaks (H1 through H4) corresponding to a set of possible reference lines (a range of possible numbers

[4] Other intonation work has claimed that a non-abstract scaling line linking the low values (also referred to as *baseline*) leads to more accurate scaling predictions (Ladd, 1993). We will not address the issue of the performance of different scaling lines here.

[5] To calculate the reference line, we are not using a model identical to Liberman & Pierrehumbert's (1984): some constants in their equation are not included here because it is unclear to us how those constants are derived.

between the f_0 minimum and the last peak). For each possible solution, we chose the reference line that produced the closest downstep ratios between peaks. The process of finding an optimal reference line could be arbitrary, in the sense that in two of the three subjects, the performance increased steadily as the value of the reference line decreased (though the difference is minimal), so the best fit was produced by the lowest value included in the calculation (i.e., a reference line set to 0). Since a reference value too close to the baseline would produce peaks on the baseline range, we chose a scaling line falling midway in between the two points (80 Hz for RS, 100 Hz for JC, and 76 Hz for AH). Those scaling lines produced rather constant downstep ratios: the mean of downstep ratios are 0.729 for RS, 0.872 for JC, and 0.871 for AH. Thus, the value of a given peak is calculated as a constant fraction (downstep ratio) of the previous one, scaling them with the reference line of each speaker (R), as follows:

$$P(x + 1) = r \times (P(x) - R) + R,$$

where $P(x)$ = peak height of a peak in position x; r = downstep factor; and R = reference line.

Below we list the linear equations (Model 2) used for our three speakers, using optimal reference and downstep values:

$$\text{Peak Height (RS)} = 0.729\ (\text{PREVPEAK} - 80) + 80,$$

$$\text{Peak Height (JC)} = 0.872\ (\text{PREVPEAK} - 100) + 100,$$

$$\text{Peak Height (AH)} = 0.871\ (\text{PREVPEAK} - 76) + 76,$$

where PREVPEAK = Previous peak height (in Hz).

Fig. 9 plots the observed mean f_0 peak values (solid lines) and compares them

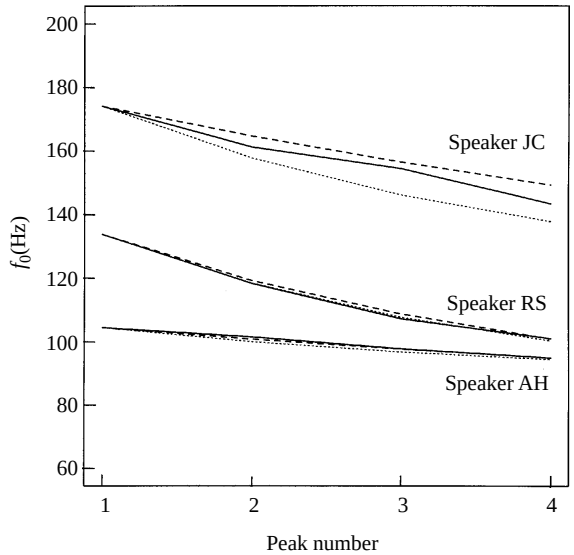

Figure 9. Observed and predicted f_0 peak values (H1 to H4). Solid lines (—) represent observed values, dotted lines (· · · ·) represent the values predicted by Model 1, and dashed lines (- - -) represent the values predicted by Model 2.

to the predicted f_0 peaks using Model 1 (dotted lines) and Model 2 (dashed lines), for each speaker. The predicted values were calculated using the mean initial peak as a starting point. As Fig. 9 shows, the values predicted by both models are quite accurate: in general, they produce contours with a sharper degree of decay between the first and second peaks and a progressively lower degree of decay over the phrase-medial highs.

3.3.5. *Comparing Models 1 and 2*

In general, the performance of the models in Sections 3.3.3. and 3.3.4. has demonstrated that the pattern of decay of peaks in Spanish downstepping contours, as in English, is successfully accounted for as a constant proportion of f_0 reduction from the previous peak. The models that take into account the f_0 values of preceding peaks (PREVPEAK) clearly supersede models that use position in the utterance (POSSENT) and/or distance between peaks (NUMSYLS) and phrasal length (NUMACC).

As Janet Pierrehumbert pointed out to us, Model 1 and Model 2 are algebraically equivalent.[6] The two models are identical and any observable difference in their fit arises because the parameter values of Model 2 (Liberman/Pierrehumbert) were probably calculated in a semi-automatic (suboptimal) fashion.

Both models (Model 1 and Model 2) are based on the same claim, namely, that the height of a given peak depends almost exclusively on the height of the preceding peak. One basic difference is the explicitness of a ''physiologically transparent'' *reference* line. The main motivation to use a reference line (Liberman & Pierrehumbert, 1984) was to constrain the peaks in longer utterances to remain above the f_0 minimum of the speaker. By including the reference line in the model, Model 2 achieves the desired result, namely, peak height in a long utterance will asymptote to this pre-set parameter. Model 1, on the other hand, achieves the same result. To show the behavior of both models on peaks in very long utterances, we produce automatically the predictions of both models on a 50-accent utterance by setting an initial value for the initial peak and letting both models do further

[6] In Model 2, the Liberman & Pierrehumbert model:

$$\mathrm{Pr} = \mathrm{P} - \mathrm{R},$$

$$\mathrm{Pr}(x+1) = r(\mathrm{Pr}(x)),$$

where $\mathrm{Pr}(x)$ = peak height (P) in position x transformed by subtracting reference line R; x = index on peak position; r = downstep factor; R = reference line.

By algebraic manipulation, given that:

$$\mathrm{Pr}(x+1) = \mathrm{P}(x+1) - \mathrm{R},$$

$$\mathrm{Pr}(x) = \mathrm{P}(x) - \mathrm{R}.$$

Then

$$\mathrm{P}(x+1) - \mathrm{R} = r(\mathrm{P}(x) - \mathrm{R})$$

$$\begin{aligned}\mathrm{P}(x+1) &= \mathrm{R} + r(\mathrm{P}(x) - r)\\ &= \mathrm{R} + r(\mathrm{P}(x)) - r\mathrm{R}\\ &= \mathrm{R} - r\mathrm{R} + r(\mathrm{P}(x))\\ &= \mathrm{R}(1-r) + r(\mathrm{P}(x))\\ &= b + a(\mathrm{P}(x)),\end{aligned}$$

where $b = \mathrm{R}(1-r)$; $a = r$.

The last line of the derivation is the equation for Model 1.

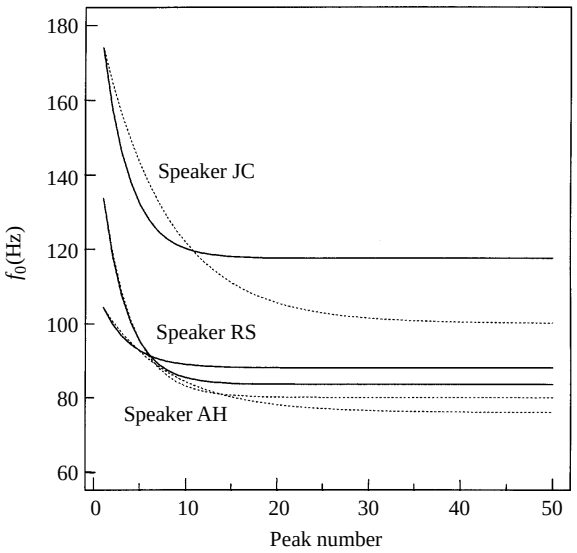

Figure 10. Predicted f_0 peak values for a given 50-accent utterance. Solid lines (—) represent the values predicted by Model 1, and dotted lines (· · · ·) represent the values predicted by Model 2.

calculations. Fig. 10 plots the values of the first 50 peaks, as predicted by both models. Model 2, as expected, asymptotes to the pre-set reference line (80 Hz for speaker RS, 100 Hz for speaker JC, and 76 Hz for speaker AH). Model 1 also asymptotes to values that are always *higher* than the constant utterance-final f_0 value for each speaker (75 Hz for speaker RS, 94 Hz for speaker JC, and 72 Hz for speaker AH).

To test the performance of both models on this data, we first extracted the matrix of pairs of observed adjacent f_0 peaks P(x) and P(x + 1)). The value of the first peak (P(x)) (for each pair) was input to both models, which produced a set of *predicted values* (P(x + 1)). Then, we computed the squared correlation coefficient (R^2, a measure of the amount of variance explained by the model) between the predicted values (P(x + 1)) of both models and the observed P(x + 1) values. Not surprisingly, Model 1 and Model 2 perfromed exactly the same for the three speakers, with differences only reflected in the decimal numbers. The squared correlation between the predicted and the observed values is 78.74% for RS, 65.79% for JC, and 64.64% for AH. This result demonstrates that the regression model (Model 1) obtains a ''hidden reference line'' automatically and produces as good a fit as Model 2.

Model 1 and Model 2, then, are expressions of the same model. The difference between the two is in ''how'' parameter values are obtained. Model 2 requires a preset reference line, which is derived from theoretical assumptions rather than from observed f_0 measurement of the pitch contour, while Model 1 is computationally convenient in the sense that all input values are taken directly from f_0 measurement. Since the two models give similar performance, we implemented Model 1 for the Spanish text-to-speech system.

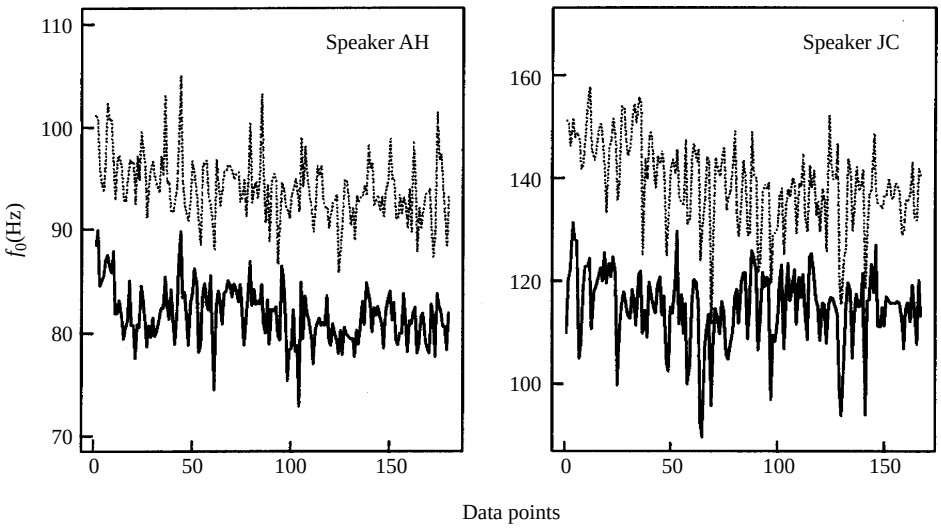

Figure 11. Observed and predicted f_0 peak values in utterance-final position for speakers AH and JC. Solid lines represent the observed values and dotted lines represent the predictions of Model 1.

3.4. *Final lowering*

Final lowering has been reported in languages like Japanese (Pierrehumbert & Beckman, 1988), Danish (Thorsen, 1981), German (Möbius, 1993: 129), and English (Liberman & Pierrehumbert, 1984: 186). This phenomenon can be described as a more drastic lowering of the pitch in the final part of an utterance, which falls far below the values predicted by the application of the downstep rule. Depending on the language, the lowering domain can be the last pitch accent (English) or a fixed time span (Japanese). Fig. 11 plots the observed f_0 values of peaks in utterance-final position (solid lines) and the predicted values using Model 1 (dotted lines). The figure clearly shows that the observed values lie consistently below the values predicted by this model.

For English, Liberman & Pierrehumbert (1984) demonstrate that utterance-final peaks can be accurately modeled by using a lowering constant, which is defined to be a fraction of the distance of P (value of the peak predicted by the downstep rule) above R (the reference line). The transformed value of the final peak is obtained by multiplying the value of the predicted peak by the final lowering constant, as follows:

$$P = R + l \times (P - R),$$

where P = peak height of the last peak in the utterance; R = reference line; l = final lowering constant ($l < 1$).

The final lowering constant for the three speakers was obtained by dividing the distance between the reference line and the observed final peak value by the reference line and the predicted downstepped peak value, as follows:

$$l = (P(obs) - R)/(P(pred) - R)$$

where P(obs) = observed peak height; P(pred) = peak height predicted by the downstep rule; R = reference line; l = final lowering constant ($l < 1$).

In the case of speaker RS, where no peak was found, we use the f_0 value at the syllable onset as the target ''peak'' value. We assume that, for a given contour, there is some choice of the final lowering constant that would put the peak right on a line between the penultimate accent and the end. The final lowering constants obtained for each speaker were the following: 0.10 for RS, 0.33 for JC, and 0.31 for AH.

To test the performance of this model, we compared its predictions to the observed results. With the final lowering constant, the performance of the model was not good: the amount of variance explained was 51% for RS, 42% for JC, and 44% for AH. As before, changing the reference line of the speaker (and the final lowering constant along with it) does not change the performance of the model. Regression analyses, using only the PREVPEAK factor, account for the same amount of variance as the previous model, with a final lowering constant: 51.2% for RS, 42.1% for JC, 44.9% for AH.

$$\text{Utt-Final Peak Height (RS)} = 25.739 + 0.692\ \text{PREVPEAK},$$

$$\text{Utt-Final Peak Height (JC)} = 34.071 + 0.716\ \text{PREVPEAK},$$

$$\text{Utt-Final Peak Height (AH)} = 41.262 + 0.42\ \text{PREVPEAK},$$

where PREVPEAK = Previous peak height (in Hz).

Still, the performance of both models is rather low (between 42 and 51% of the variance is explained). The use of a constant ratio of decay from the previous peak (or the predicted peak) has the desirable effect of capturing the fact that final peaks are subjected to less lowering in longer utterances than in shorter utterances. Consistently, a final lowering constant produces a larger amount of lowering in shorter utterances than in longer ones since it applies to accents with a higher f_0 value in the former case. We observe a tendency of the model to predict both *undershooting* and *overshooting* in phrase-final peaks in different phrasal positions: in shorter utterances, the predicted final peaks are too low, while in longer utterances, the predicted peaks are too high. This is due to a tendency of final peaks to level out at a given point in the utterance: the *final lowering* ratios of two, three, four, and five pitch accents have a clear tendency to decrease for the three speakers. The three plots in Fig. 12 show the values predicted by the model (using the model with the final lowering constant) as a function of the observed values for different utterance lengths (2: two accents; 3: three accents; 4: four accents; and 5: five accents). The solid line is $x = y$, where the medial values are the same as the observed values. The difference between the dotted lines and the solid lines highlights a difference between the behavior of the observed and predicted values: consistently, predicted values in shorter utterances (2) are lower than observed values, and in longer utterances (5) they are higher. Thus, a lowering *constant* has the undesirable effect of predicting the same amount of decay, regardless of the utterance position. Adding a phrasal length factor to the linear model above substantially increases the performance of the model: 54.5% for RS, 49.7% for JC, 47.1% for AH.

3.4.1. *Final fall*

The goal of this section is to test the behavior of another possible account of RS's *final fall*, what we will call the *interpolation* model. Evidence about the behavior of the f_0 values at the onset and offset of the final accented syllables provided

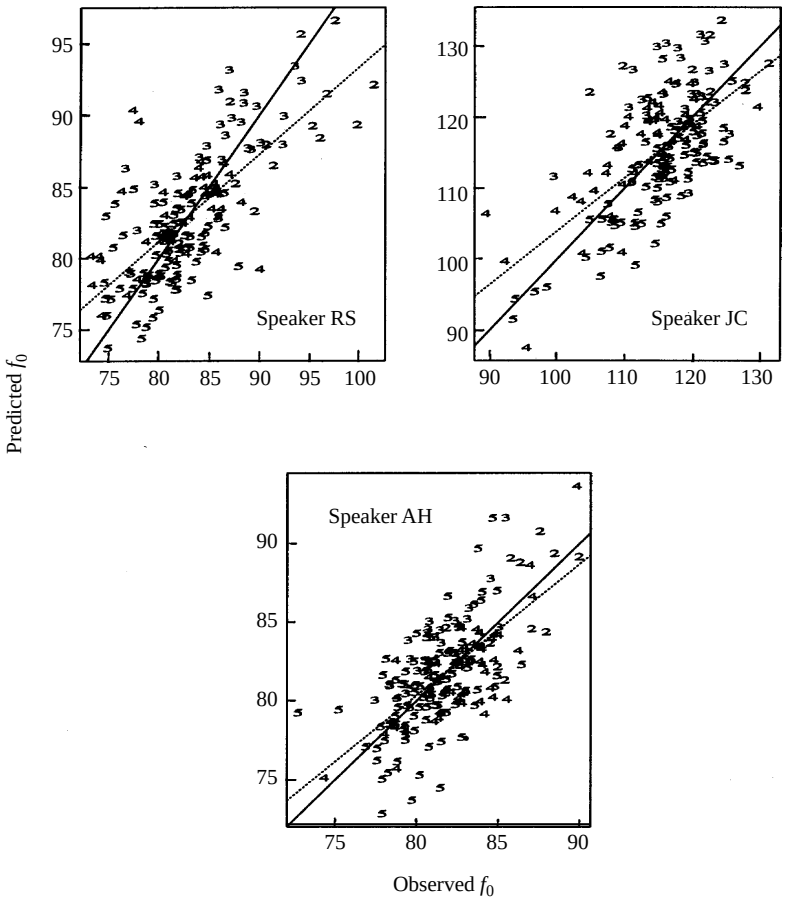

Figure 12. Predicted f_0 peak values in utterance-final position (y-axis) as a function of observed f_0 peak values (x-axis) for the three speakers. The numbers in the plots refer to utterance length (the number of accents). The solid line is a reference line showing where x = y, and the dotted line is a regression line representing the correlation between the observed and predicted f_0 values.

preliminary evidence that RS's *final fall* could be better characterized as an interpolation line linking the peak of the penultimate pitch accent and the final near-constant f_0 value.

Fig. 13 shows four mean schematized final contours of RS in phrases of different lengths (from two to five pitch accents). The three points selected correspond to the onset of the vowel bearing the phrase-final accent (final H in our labeling scheme), its offset (penultimate L), and the phrase-final f_0 values (final L), respectively. First, the data in Fig. 13 reflect the different behavior of the pitch contour over the final accent in a sequence, which always displays a continuously decreasing pitch level rather than a distinct pitch peak. Moreover, f_0 values taken at the onset and offset are negatively correlated with phrasal length: f_0 values at both vowel onset and offset tend to decrease as phrasal length increases. Even though not all differences are statistically significant (see standard error bars in Fig. 11), the effect is quite

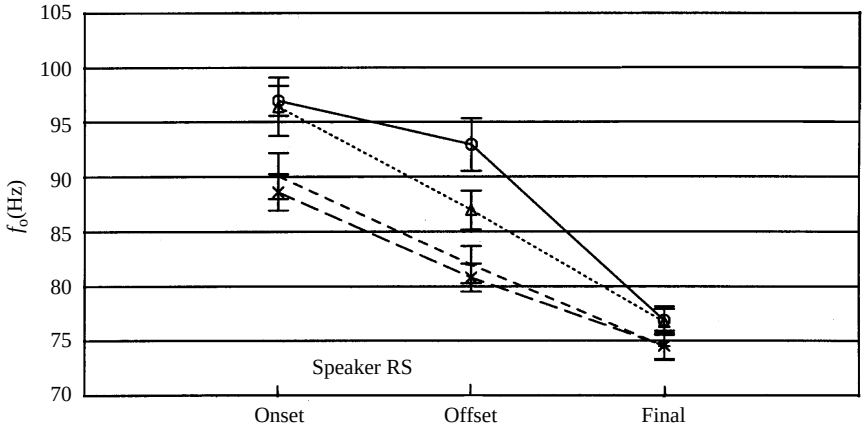

Figure 13. Mean schematized *final falls* corresponding to utterances of different lengths (from two to five pitch accents). The three values along the x-axis correspond to f_0 values at the onset of the last accented vowel, at its offset, and at the end of the utterance, respectively. The height of the bars represents standand errors. ——, Two accents; · · · ·, three accents; - - -, four accents; — —, five accents.

consistent. This effect is not surprising if we take into account that longer utterances have accumulated more downstep effect (i.e., the further into the utterance the pitch accent occurs, the lower its peak).

Is there any evidence of lower pitch values at the onset and offset points when there are more intervening unstressed syllables? Fig. 14 displays six mean schematized *final falls* in phrases of different lengths (from three to five accents), plotting

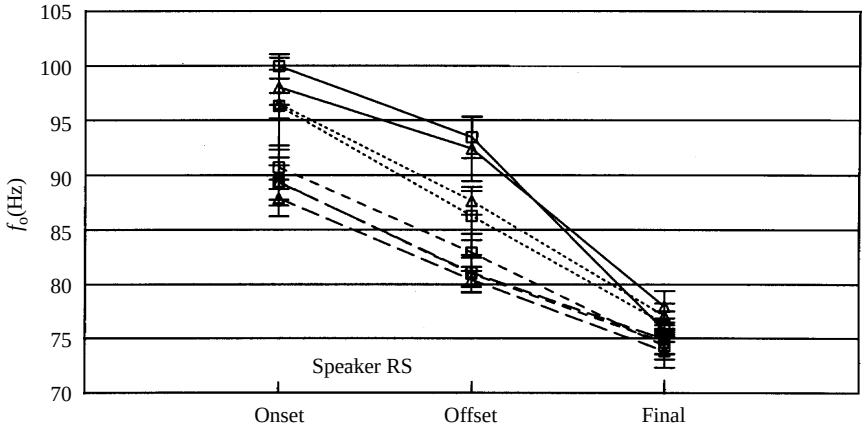

Figure 14. Mean schematized *final falls* corresponding to utterances of different lengths (from two to five pitch accents) and with different numbers of unstressed syllables before the accent. The three points along the x-axis correspond to f_0 values at the onset of the last accented vowel, its offset, and at the end of the utterance, respectively. "□" represents points in utterances with two preceding unstressed syllables and "△" represents those with three unstressed syllables. ——, Two accents; · · · ·, three accents; - - -, four accents; — —, five accents.

separately utterances with two or three unstressed syllables preceding the last accented vowel (marked by distinct plotting symbols). Even though not entirely consistent, the pitch level at the onset and offset positions tends to be lower when the distance in time to the previous accent increases (the symbol "△" is generally placed lower than the symbol "□" in utterances of the same length).

Thus, preliminary evidence seems to indicate that the straight-line model can predict RS's *final fall* better than a final lowering constant. First, there is no evidence for the presence of target values for the last accented syllable (onset or offset), as they both progressively decrease as phrasal length increases. Second, the steepness of the terminal fall depends on both utterance length and the number of syllables composing the last part of the utterance. Similarly, the final slope is progressively less steep both as the utterance becomes longer (from two to five pitch accents) and as the number of unstressed syllables between the final two pitch accents increases.

To make an explicit comparison between the two models (the interpolation model and the linear model), we compared the predictions of both models by taking as input pairs of observed f_0 values for speaker RS. Yet, the amount of variance explained by the interpolation model is relatively low, 39.60%. All the models in the previous section performed better (between 45 and 53%).

$$\mathrm{P(Hz)} = \mathrm{PrevPeak(Hz)} - (\mathrm{PrevPeak(Hz)} - \mathrm{Fin(Hz)}/(\mathrm{PrevPeak(t)} - \mathrm{Fin(t)}) * (\mathrm{PrevPeak(t)} - \mathrm{P(t)}),$$

where P(Hz) = f_0 value at point x in the final fall; P(t) = time value at point x in the final fall; PrevPeak(Hz) = f_0 value of the previous peak in utterance (Hz); PrevPeak(t) = time of the previous peak in utterance (ms); Fin(Hz) = f_0 value at utterance-final position; Fin(t) = time value at utterance-final position.

An explicit regression analysis comparison between the straight-line treatment of RS's final pitch accent and the optimal fit using a final lowering factor has shown that models that assume a constant ratio of decay lead to an objectively better fit of the data. Thus, RS's *final fall* is shown to behave as other final peaks; even though on the surface we find no stable f_0 target peaks in RS's contours, we found that models which assume a target value in the onset work better. Thus, we claim that the two possible realizations of the final accent (namely, peak *vs.* falling contour) involve the same phonological option (H* accent), and the only difference between the two is the phonetic implementation of the two accents.

4. Conclusion

The study of f_0 peak height in Spanish downstepping contours has revealed that the observed downtrend patterns can be largely explained by a linguistically controlled *downstep*. Regression analyses demonstrated that the main predictor of peak height in our data was *height of the previous peak.* As in English (Liberman & Pierrehumbert, 1984), the H downtrend pattern was accurately modeled as an exponential decay to a constant nonzero asymptote (between 55 and 65% of the variance of the data could be accounted for). Applied iteratively, this rule produces the peak pattern observed in Spanish, with a more drastic decay between the first

and second peaks and less lowering between subsequent peaks. An explicit comparison between models that use previous peak height *vs*. phrasal position as the main predictor indicates that models that predict peak height as a ratio of the preceding peak perform better. Moreover, models that enforce scaling lines in the linear equation (like the *reference line* in Pierrehumbert & Liberman's case) are shown to be equivalent to other models that do not compute directly such scaling lines.

Effects of factors such as utterance length and temporal distance between peaks were found to be small and non-significant across speakers. Values of f_0 peaks in the same phrasal position had a tendency to reach similar levels, regardless of factors such as phrasal length or distance between accents. Still, since controlled duration intervals between peaks were relatively small in our study (between 100–150 ms), stronger confirmation of the lack of declination in downstepping contours would be needed by using a database with accents separated by more than three syllables. However, independent observation that clash situations trigger the same amount of peak lowering in Spanish downstepping patterns (Prieto & Shih, 1995) seems to independently confirm the lack of a time-dependency effect.

Utterance-final peaks in Spanish undergo a *final lowering* phenomenon, as has been described for other languages. That is, the amount of lowering between the last two peaks of an utterance is consistently lower than expected by using the speaker-dependent *downstep* ratio. Our speakers produced final accents in two different ways, namely, as compressed final peaks and as continuously falling slopes. It is shown that both shapes are best fitted by assuming the presence of a target at a given point in the stressed syllable (at the onset of the stressed syllable for the falling slope shape). Similar to the downstep rule, this target is reasonably fitted by assuming a constant decay from the preceding peak (or the peak predicted by the downstep rule). A statistical comparison between the performance of two models shows that models that assume an optimal choice of a final *lowering constant* (which applies to the output of the downstep rule) are equivalent to models that assume a particular ratio of decay (larger than the downstep ratio) between the last two peaks of the utterance. Final peaks in our data were shown to behave significantly different from previous peaks in the following respect: unlike utterance-medial Hs, utterance-final Hs displayed a tendency to maintain relatively high f_0 levels, or at least higher than predicted by the models; those values consistently *overshot* the predicted final peak values, which could reflect a constraint on how low peaks can fall. Indeed, in this case, adding the phrasal length factor in the model increased substantially the fit of the final peak data.

In light of these results, we argue that the descending H patterns in Spanish downstepping contours are almost exclusively accounted for by the repeated application of a downstep rule. In general, this result supports the idea that f_0 contours can be decomposed into series of high target points that are controlled at a local level, in agreement with Liberman & Pierrehumbert (1984) and Pierrehumbert (1980).

We want to thank Cinzia Avesani, Julia Hirschberg, José Ignacio Hualde, Bernd Möbius, Jennifer Venditti, and especially Jan van Santen for very helpful discussion and advice at different stages of the writing of this paper. We are also thankful to our Mexican speakers, José Cervantes, Arturo Hale, and Rodney Sidransky for their patience and collaboration in

the recording sessions, and to Pere Gifra, Eduardo Lleida, José Pérez, and Joan Salavedra, who participated in the perceptual tests. Finally, we would like to thank the editor of our paper, Janet Pierrehumbert, and two anonymous reviewers, for insightful comments that lead to substantial improvement of this paper..

.

References

Bruce, G. (1977) *Swedish word accents in sentence perspective.* Gleerup: Lund.

Cooper, W. F. & Sorensen, J. M. (1981) *Fundamental frequency in sentence production.* Heidelberg: Springer.

Delattre, P., Olsen, C. & Poenack, E. (1962) A comparative study of declarative intonation in American English and Spanish, *Hispania,* **45**, 233–241.

Fant, L. (1984) *Estructura informativa en español. Estudio sintáctico y entonativo.* Acta Universitatis Upsaliensis 34. Uppsala.

Fujisaki, H. (1983) Dynamic characteristics of voice fundamental frequency of speech and singing. In *The production of speech* (P. F. MacNeilage, editor), pp. 39–55. New York and Berlin: Springer-Verlag.

Fujisaki, H. (1988) A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In *Vocal physiology: voice production, mechanisms and functions* (O. Fujimura, editor), pp. 347–355. New York: Raven.

Ladd, D. R. & Johnson, C. (1987) ''Metrical'' factors in the scaling of sentence-initial accent peaks, *Phonetica,* **44**, 238–245.

Ladd, D. R. (1993) On the theoretical status of ''the baseline'' in modelling intonation, *Language and Speech,* **36**, 435–451.

Liberman, M. (1975) The intonational system of English. Ph.D. dissertation, MIT. Cambridge, MA.

Liberman, M. & Pierrehumbert, J. (1984) Intonational invariance under changes in pitch range and length. In *Language sound structure* (M. Aronoff & R. Oehrle, editors), pp. 157–233. Cambridge, MA: MIT Press.

Lieberman, P. (1967) *Intonation, perception, and language.* Cambridge, MA: MIT Press.

Möbius, B. (1993) *Ein quantitatives Modell der deutschen Intonation. Analyse und Synthese von Grundfrequenzverlaeufen.* Tübingen: Niemeyer.

Navarro-Tomás, T. (1944) *Manual de entonación española.* New York: Hispanic Institute in the United States.

Pierrehumbert, J. (1979) The perception of fundamental frequency declination, *Journal of Acoustics Society of America,* **66**, 363–369.

Pierrehumbert, J. (1980) *The phonology and phonetics of English intonation.* Ph.D. Dissertation, MIT.

Pierrehumbert, J. & Beckman, M. (1988) *Japanese tone structure.* Cambridge, MA: MIT Press.

Pitrelli, J., Beckman, M. & Hirschberg, J. (1994) Evaluation of prosodic transcription labeling reliability in the ToBI framework. *Proceedings of the third international conference on spoken language processing,* **2**, 123–126, Yokohama, Japan.

Poser, W. J. (1984) The phonetics and phonology of tone and intonation in Japanese. Ph.D. Dissertation, MIT.

Prieto, P., van Santen, J. & Hirschberg, J. (1995) Tonal Alignment Patterns in Spanish, *Journal of Phonetics,* **23**, 429–451.

Prieto, P. & Shih, C. (1995) Effects of tonal clash on downstepped H* accents in Spanish. *Proceedings of EUROSPEECH '95: fourth European conference on speech communication and technology,* volume 2, pp. 1307–1310, Madrid, Spain.

Prieto, P., Nibert, H. & Shih, C. (1996) The absence or presence of a delination effect on the descent of f_0 peaks?: evidence from Mexican Spanish. In *Grammatical theory and Romance languages* (Karen Zagona, editor), pp. 197–207. Philadelphia: John Benjamins Publishing.

Prieto, P. Register shift in Spanish downstepping contours. Article in preparation.

Sosa, J. M. (1991) Fonética y fonología de la entonación del español hispanoamericano. Ph.D. Dissertation, University of Massachusetts at Amherst.

Sternberg, S., Wright, C. E., Knoll, R. L. & Monsell, S. (1980) Motor programs in rapid speech: additional evidence. In *Perception and production in fluent speech* (R. A. Cole, editor), pp. 507–534. Hillsdale, NJ: Lawrence Erlbaum Associates.

Talkin, D. (1989) Looking at speech, *Speech Technology,* **11**, 384–400.

Thorsen, N. (1980) Intonation contours and stress group patterns in declarative sentences of varying length in ASC Danish, *Annual report of the Institute of Phonetics,* University of Copenhagen, **14**, 1–29.

Thorsen, N. (1981) Intonation contours and stress group patterns in declarative sentences of varying length in ASC Danish—supplementary data, *Annual report of the Institute of Phonetics,* University of Copenhagen, **15**, 13–47.

van den Berg, R., Gussenhoven, C. & Rietveld, T. (1992) Downstep in Dutch: implications for a model. In *Papers in laboratory phonology II: segment, gestures, tone* (G. J. Docherty & D. R. Ladd, editors), pp. 335–367. Cambridge: Cambridge University Press.

Appendix 1

Set of target phrases used in the production experiment

TWO ACCENTS
1: ráyo de lúna
2: ráyo de la lúna

THREE ACCENTS
3: ráyo de lúna de máyo
4: ráyo de lúna de mi máyo
5: ráyo de mi lúna de máyo
6: ráyo de mi lúna de mi máyo

FOUR ACCENTS
7: ráyo de lúna de máyo de gála
8: ráyo de lúna de máyo de la gála
9: ráyo de lúna de mi máyo de gála
10: ráyo de lúna de mi máyo de la gála
11: ráyo de la lúna de máyo de gála
12: ráyo de la lúna de máyo de la gála
13: ráyo de la lúna de mi máyo de gála
14: ráyo de la lúna de mi máyo de la gála

FIVE ACCENTS
15: ráyo de lúna de máyo de gála de Lóla
16: ráyo de lúna de máyo de gála de la Lóla
17: ráyo de lúna de máyo de la gála de Lóla
18: ráyo de lúna de máyo de la gála de la Lóla
19: ráyo de lúna de mi máyo de gála de Lóla
20: ráyo de lúna de mi máyo de gála de la Lóla
21: ráyo de lúna de mi máyo de la gála de Lóla
22: ráyo de lúna de mi máyo de la gála de la Lóla
23: ráyo de la lúna de máyo de gála de Lóla
24: ráyo de la lúna de máyo de gála de la Lóla
25: ráyo de la lúna de máyo de la gála de Lóla
26: ráyo de la lúna de máyo de la gála de la Lóla
27: ráyo de la lúna de mi máyo de gála de Lóla
28: ráyo de la lúna de mi máyo de gála de la Lóla
29: ráyo de la lúna de mi máyo de la gála de Lóla
30: ráyo de la lúna de mi máyo de la gála de la Lóla

Appendix 2

Set of target phrases used in the perceptual test.

Ráyo de lúna
Ráyo de la lúna
Ráyo de mi lúna de mi máyo

Ráyo de lúna de máyo
Ráyo de mi lúna de máyo
Ráyo de la lúna de máyo de gála
Ráyo de lúna de mi máyo de la gála
Ráyo de la lúna de máyo de la gála
Ráyo de lúna de máyo de gála
Ráyo de lúna de mi máyo de gála de Lóla
Ráyo de la lúna de mi máyo de la gála de Lóla
Ráyo de lúna de mi máyo de gála de la Lóla