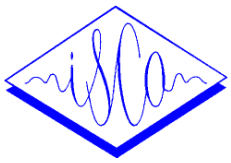

Second ESCA/IEEE Workshop
on Speech Synthesis
Mohonk, New Paltz, NY, USA
September 12-15, 1994

PATTERNS OF F0 PEAK PLACEMENT IN MEXICAN SPANISH

Pilar Prieto
Jan van Santen
Julia Hirschberg
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974–0636, USA
{prieto, jphvs, julia}@research.att.com

Abstract

A speaker of Mexican Spanish read 405 declarative sentences containing nine distinct target syllables with an H* accent, under different prosodic conditions: end of intonational phrase, end of intermediate phrase, and phrase-medial (varying syllabic position in the word and distance to the next stressed syllable). A preliminary analysis shows that intrasyllabic segmental durations and prosodic factors such as adjacency to word, intonational, and intermediate boundaries, as well as stress clash, are key components in the prediction of peak location.

Introduction

To develop an f0 assignment algorithm for a Mexican Spanish text-to-speech system, we designed an experiment to study the f0 peak alignment patterns of declarative sentences in this dialect. The main motivation for the present study was the observation that phrase-medial f0 peaks of H* accents (simple peaks, as described in Pierrehumbert 1980) in Mexican Spanish were usually displaced to the right of the accented syllable perceptually associated with the high tone, but that this displacement was not uniform throughout. Although this phenomenon has been noticed by Navarro-Tomás (1944) and Fant (1984) for Peninsular Spanish, no attempt has been made to identify the factors that influence H* peak timing in Spanish. Some perceptual experiments with the implementation of synthetic Spanish declarative contours (Martí 1992) indicate that the modelling of peak placement is a perceptually important part of Spanish f0 assignment algorithms.

Experimental Design and Data Preparation

Timing of fundamental frequency peaks in read declarative sentences was studied as a function of the following linguistic factors:

1) duration of the segments comprising the accented syllable; 2) degree of prominence of the H* target accent; 3) distance (in syllables) from the accent to intermediate and intonational phrase boundaries; 4) distance (in syllables) from the accent to the end of the word; 5) distance (in syllables) from the accent to the next stressed syllable.

The database consisted of three types of sentences,[1] designed to trigger increasing degrees of prominence in the target H* pitch accent. In order to minimize pitch tracking errors, target words with only sonorant consonants were used in the experiment.[2] The target words were placed in the following phrasal positions for the three sentence-types: a) end of intonational phrase; b) end of intermediate phrase; and c) sentence-medial position. Each of the conditions consistently varied position of the accentable syllable in the word and distance to the next stressed syllable.[3]

[1]Type 1-3 sentences are the following (the accentable syllable is in boldface):

1. *Se suponía que tenía que murmurar numero, pero murmuró* **número**. 'He was supposed to murmur "I number", but he murmured number.'
2. *El no murmuró húmedo. Murmuró* **número**. 'He did not murmur humid. He murmured number.'
3. A: *Murmuró la palabra secreta?. B: Pues, murmuró la palabra* **número**. *Es ésta la palabra secreta?* 'A: Did he murmur the secret word? B: Well, he murmured the word "I number". Is that the secret word?.'

[2]We illustrate the nine words selected, and the symbols used to code within-word position:

Initial=I	Medial=M	Final=F
número	numero	numeró
nómina	nomina	nominó
lámina	lamina	laminó

[3]We use the following notational scheme, which we illustrate with examples for Type 2 sentence with the words *número, numeró*. End of an intonational phrase: END-INTON[i]; end of intermediate phrase: END-INTER[i]; and

A male adult Mexican speaker[4] read the test sentences three times at a normal speech rate (for a total of 405 utterances). The recorded sentences were inspected, prosodically labeled and segmented.

Analysis

Effects of Syllable Duration

The data show a high correlation between peak delay (distance from the f0 peak to the onset of the stressed syllable) and duration of the stressed syllable, in all conditions defined by the factors (correlation coefficients range from 0.34 to 0.85, all significant at p=0.05 or better). Figure 1 plots peak delay as a function of syllable duration in the groups MED[F,I] (*numeró rápido*) and MED[I,I] (*número rápido*).[5]

Also note that in the MED[I,I] condition, the syllable peak is located roughly equally often on either side of the syllable boundary, whereas in the MED[F,I] condition the peak is always located in the stressed syllable.

The strong correlations indicate that peaks are *not* located at some fixed distance to the onset of the syllable, but shift in rightward direction as syllable duration increases. The shift is not proportional, because this would imply that the relative contribution of the onset and vowel durations to this shift should be the same (multiple regression analyses on the onset and vowel duration variables give mean regression coefficients of

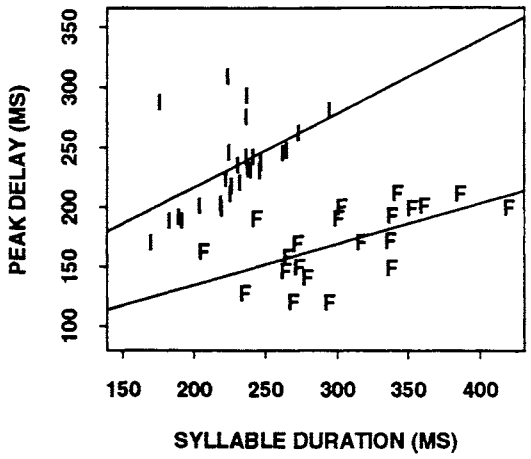

Figure 1: Peak delay (relative to the onset of the stressed syllable), as a function of syllable duration in two conditions: MED[F,I] —*numeró rápido*— and MED[I,I] —*número rápido*—.

0.716 and 0.395, respectively).[6] In other words, rather than representing peak delay as a proportion on the entire syllable, it is more accurately represented as the following weighted combination of the durations of the onset and the vowel:

Peak Delay = 0.716 Onset Dur + 0.395 Vowel Dur

Effects of peak prominence

In the data recorded, sentences of Types 1-3 consistently elicited three distinct degrees of prominence of the H* pitch accent (mean values of absolute f0 maximum in the condition END-INTON[I] are: Type 1: 128.97 Hz; Type 2: 145.95 Hz; Type 3: 186.79 Hz). Yet, against our expectation, peak delay values did not increase with higher peaks (either comparing the average values in the three sentence types, or computing correlations within-sentence types).

Effects of intonational and intermediate boundaries

Figure 2 plots the mean values of peak delay as a function of within-word position (I,M,F) for the three phrase boundary conditions: intonational phrase

sentence-medial position: MED[i,j]. Here, i = stress location of the target word [Initial: I, Medial: M, and Final: F]; j = stress location of the word following the target word [Initial: I, Medial: M, and Final: F]. Below, the target accented syllable is in boldface:

1. END-INTON[I]
El no murmuró húmedo. Murmuró **núm**ero.
2. END-INTER[F]
El no murmuró húmedo. Murmuró **numer**ó, estoy seguro.
3. MED
3.1. MED[I,I]
El no murmuró húmedo. Murmuró **núm**ero rápido.
3.2. MED[I,M]
El no murmuró húmedo. Murmuró **núm**ero nervioso.
3.3. MED[I,F]
El no murmuró húmedo. Murmuró **núm**ero regular.
3.4. MED[F,I]
El no murmuró húmedo. Murmuró **numer**ó rápido.

We use the notation MED[,j] to refer to the average of MED[I,j], MED[M,j], and MED[F,j]; likewise for MED[i,]. MED[F,I], MED[M,I], and MED[I,I]).

[4]RS, our speaker, is a native of the city of Ciudad Juárez, Chihuahua, in the North of Mexico.

[5]See Footnote 3 for a description of the notational conventions.

[6]As in English (van Santen & Hirschberg 1994), onset duration regression coefficients are consistently larger than vowel duration coefficients.

boundary, intermediate phrase boundary, and phrase-medial condition (confining the analysis to MED[,M]).

First, the figure shows an overall effect of intonational phrase boundaries (peak delay is lower than in other phrasal conditions). Since the same target syllables were recorded for the intermediate and intonational boundary groups, we could perform a paired t-test analysis. All pairwise tests comparing the two groups for each of the three within-word positions (I, M, F) were statistically significant (at $p < 0.02$).

The difference between intermediate and intonational boundaries might be related to the fact that f0 phrase-final values are 10 Hz lower at the end of intonational boundaries.[7] We found no significant differences in syllable duration that can explain these peak location differences.

The figure (MED[,M] condition) also shows a clear effect of within-word position regardless of whether there is a following phrase boundary. Importantly, within-word position has much stronger effects when there is such a boundary (about 90 ms, comparing M with F) than when there is no such boundary (about 25 ms). Thus, location is affected both by within-word position and by proximity to the phrase boundary.

Stress-clash effects

To investigate the long-range effects of stress clash,[8] we encoded both the number of subsequent unstressed syllables (0,1,2,3,4) and within-word position (I,M,F). Figure 3 plots mean peak delay, relative peak delay, syllable-duration, and peak-to-end values as a function of the distance to the next stressed syllable. The plots show that the strict stress clash condition (*adjacency*: number of subsequent unstressed syllables = 0) indeed produces the most retraction, no matter how peak location is measured. However, stress clash also appears to have effects at a longer range: There is a weak but consistent tendency to increase peak delay as the number of unstressed syllables increases.

A Linear Model of Peak Delay

We were able to capture the joint effects of the factors by means of linear regression, using slightly different models for the three location types:[9]

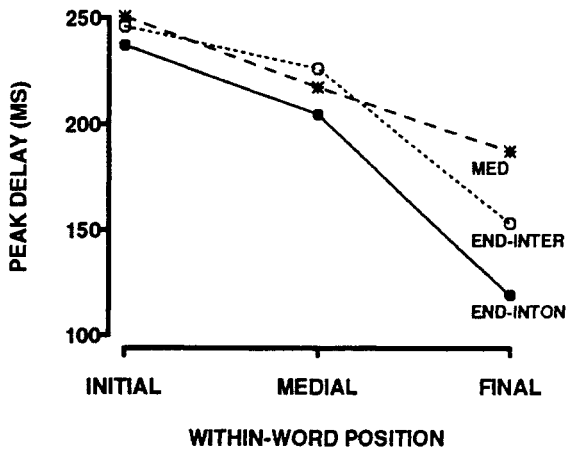

Figure 2: Mean values of peak delay in syllables before intonational (END-INTON) or intermediate phrase boundaries (END-INTER), and in phrase-medial position (MED[,M]), varying within-word position (Initial, Medial and Final).

Location Type	Factors Included			
END-INTON	*OD*	*VD*	*DEP*	
END-INTER	*OD*	*VD*	*DEP*	
MEDIAL	*OD*	*VD*	*DEW*	*DNS*

The regression analyses explained between 65% to 80% of the variance. Onset duration and vowel duration regression weights were consistent with the weights reported in a previous section. The other three factors (distance to word or phrasal boundaries, and distance to next stressed syllable), were analyzed in two ways, namely, encoding a binary or a non-binary distinction (i.e., the first encoded *adjacency*, the second *proximity* in syllables). In general, keeping the gradual distinctions in the number of syllables improved the prediction rate slightly.

The regression coefficients of all factors included in the analysis were statistically significant. Excluding factors such as distance to word and phrasal boundaries, and to next stressed syllables decreased the performance of the model,[10] showing that such prosodic factors are key components in the prediction of peak

[7]For example, the planning of a final contour with lower final values could have the effect of pushing back the peaks.

[8]That is, effects of the distance to an upcoming stressed syllable.

[9]*OD* = Onset Duration; *VD* = Vowel Duration; *DEP* = Distance to End of Phrase; *DEW* = Distance to End of Word; *DNS* = Distance to Next Stressed Syllable. For the last three factors, distance is computed in syllables. Peak prominence was not included in these analyses.

[10]The percentage variance explained decreased to 45% in the MEDIAL case, to 66% in the END-INTON case, and to 34% in the END-INTER case.

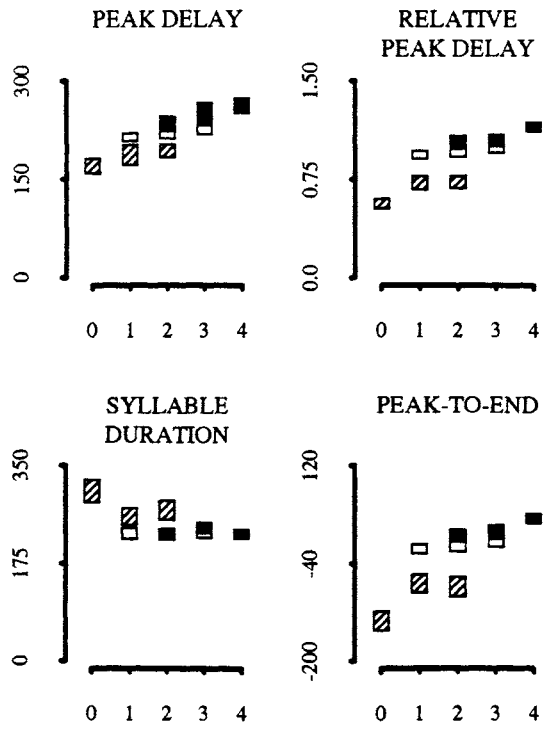

Figure 3: Mean values of various measures of peak delay – and of overall syllable duration – in Initial (black rectangles), Medial (white rectangles), and Final within-word position (striped rectangles), at an increasing distance (in syllables) to the next stressed syllable (horizontal axis). *Relative peak delay* is peak delay divided by syllable duration, and *peak-to-end* is peak delay minus syllable duration. The height of the rectangles represent 95% confidence intervals.

location.

Finally, we performed the same analysis expressing peak timing as a proportional value of the syllable, and the performance decreased (mean rms deviation values are smaller for the absolute distance case: 0.028 seconds vs. 0.030 seconds).

Conclusion

The data from the present experiment show that onset/rime duration and right prosodic context (in particular, adjacency to a word-boundary, type of phrasal boundary and stress clash) are key factors in the prediction of the f0 peak location of H* accents in Mexican Spanish. In general, as vowel and onset durations increase, peak delays also increase. Yet, when vowels are lengthened by upcoming prosodic units, peak delay values decrease.

Taking into account these factors, we have used a regression model that predicted a significant portion of the variance of peak placement. Better prediction was achieved using raw peak delay rather than relative peak delay.

In light of these results, it seems that we can safely conclude that at least two parallel structures affect timing control of f0 contours. On the one hand, the durations of the segments making up the target accented syllable seem to determine peak delay. In general, the longer the segments are, the longer the rise time is. On the other hand, prosodic units such as word and phrasal boundaries have retracting effects on peak placement. In some cases, the reason for the peak being pushed back could be attributed to a gestural need to accommodate an upcoming f0 realization (for example, the accommodation of a following low tone at the end of intermediate and intonational phrases, or an upcoming rising accent in tonal clash cases). Yet, in other cases, f0 timing seems not to involve gestural overlap with other tonal specifications, but phonological preplanning. Some evidence for this phonological preplanning is the fact that f0 peaks tend to respect word-boundaries, even though there is no clear tonal specification that needs to be accommodated either before or after.

References

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

Martí, Jordi. 1992. *Modelització suprasegmental i processament lingüístic per a un conversor text-veu.* Master's Thesis. Universitat Politècnica de Catalunya.

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

Pierrehumbert, Janet. 1980. *The Phonology and Phonetics of English Intonation.* Ph.D. Dissertation, MIT.

Silverman, Kim E. A. & Pierrehumbert, Janet. 1990. The timing of prenuclear high accents in English. In Kingston, John & Mary Beckman (eds) *Papers in Laboratory Phonology.* Cambridge University Press: Cambridge, MA.

Steele, S. A. & Altom, M. J.. 1986. Nuclear Stress: Changes in Vowel Duration and F0 Peak Location. TM 11227-860625-09. AT&T Bell Laboratories.

van Santen, Jan P. H. & Hirschberg, Julia. 1994. Segmental Effects on Timing and Height of Pitch Contours. *1994 Proceedings of the International Conference on Spoken Language Processing, Japan.*