INVESTIGATIONS OF F0 CONTROL: PITCH TARGETS VS. PITCH REGISTER

Seung-Eun Kim[1], Sam Tilsen[2]

[1]Northwestern University, [2]Cornell University
seungeun.kim@northwestern.edu, tilsen@cornell.edu

ABSTRACT

This study examines speakers' control of F0 by evaluating *target*- and *register*-control hypotheses. The *target*-control hypothesis holds that speakers adjust individual pitch targets to produce variations in F0, while the *register*-control hypothesis holds that speakers adjust pitch register (in which the pitch targets are defined) to vary F0. These hypotheses are evaluated with empirical F0 trajectories, examining correlations between F0 peaks and valleys. The results found that the peaks and valleys are not independently controlled, and that they are positively correlated. This suggests that speakers may control F0 by adjusting pitch register rather than by changing targets, and with a greater extent of manipulation of register ceiling or floor compared to span.

Keywords: F0 control, pitch targets, pitch register, correlation

1. INTRODUCTION

This study is motivated by the insight that there are two ways in which the control of F0 can be accomplished. One is that the speakers adjust individual *pitch targets*, and the other is that they adjust *pitch register*, understood here as a sensorimotor representation of the F0 control space in which the targets are defined. We refer to the first possibility as the *target*-control hypothesis and the second as the *register*-control hypothesis, and we aim to assess which one better accounts for empirical patterns in F0 trajectories over multi-phrase utterances.

A pitch target is defined here as a cognitive representation of F0 that speakers want to achieve while speaking. Speakers are assumed to implement a series of F0 goals or targets during production, and the vocal control system has parameters that determine these values. Theories of intonation have conceptualized the notion of pitch target in various ways. For instance, in the Autosegmental-Metrical intonation phonology (e.g. [1], [2], [3], [4]), speakers are considered to aim for distinctive pitch levels, and these targets are realized as peaks and valleys in surface F0 contours. The PENTA model proposed by [5] holds pitch targets to be either levels or rises/falls, which however are the underlying targets that do not necessarily map to the surface peaks and valleys. In some computational F0 models (e.g. [6], [7], [8]), pitch targets are defined with parameters that specify target values, forms/shapes, and durations.

Pitch register is commonly referred to as the range of F0 values (F0 space) that speakers can produce and utilize at a given time in an utterance. It is defined by a combination of at least two of the following three parameters: ceiling, floor, and span. According to [4], the notion of pitch register has been introduced in the F0 literature under different names, such as "tonal space" [9] in resemblance to vowel space, "tonal level frame" [10], "transform space" [3], or "grid" [11]. While all these terms represent the notion of space or range, some models lack a full specification of range, and instead specify a rough position within the range that pitch targets are superimposed on (e.g. [7], [8]).

Figure 1 illustrates the two different (yet not mutually exclusive) control strategies. Under the (i) *target*-control hypothesis, variation in the values of the F0 peaks and valleys within an utterance is directly manipulated by speakers. As shown in Figure 1-(i), to produce an F0 contour with two different F0 peaks, speakers would have two distinct high (H) pitch targets in mind, for example one at the top of the current pitch range (1.0) and the other at about 60% of the range (0.6). In contrast, under the (ii) *register*-control hypothesis, F0 variation arises from changes in register: speakers adjust the tonal space, while the targets remain constant. This is reflected in Figure 1-(ii): although targets are same across prosodic units (both are 1.0), the register shift results in different surface F0 peaks.

Evidence for these hypotheses is evaluated using F0 trajectories extracted from the experimental data. The experiment elicited utterances with one, two, or three subject noun phrases (NPs). The F0 contours

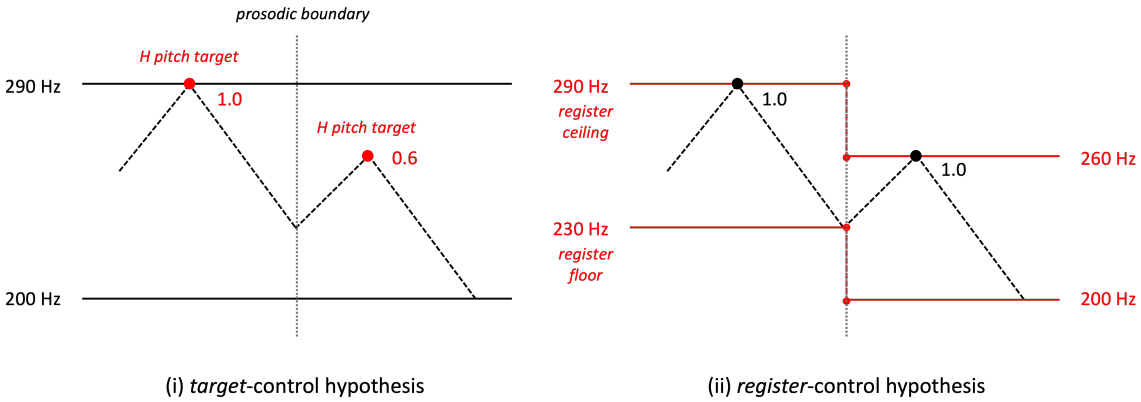

Figure 1: Schematic illustrations of F0 control hypotheses. The black dashed line represents a schematic F0 contour, the solid horizontal lines show pitch register ceiling and floor, and the dots at the peaks indicate high (H) pitch targets. A hypothetical prosodic boundary (prosodic word or phrase) is indicated as a vertical dotted line. The main F0 parameter that leads to variations in F0 under each hypothesis is marked in red. Arbitrary parameter values are provided for H targets and register ceiling/floor.

that we analyze had an F0 valley, a peak, and another valley at each NP. This may be represented in the Autosegmental-Metrical framework as an L+H*-L sequence, although nothing in our analysis depends on the validity of this representation.

Since the control parameters cannot be observed directly, F0 peaks and valleys are assumed to reflect high (H) and low (L) pitch targets, and F0 ranges to reflect register span. Although surface F0 peaks, valleys, and ranges may not exactly reflect internal representations of pitch targets or register, these surface measures are likely to be highly correlated with the underlying control parameters. We note that this assumption is most tenable in the moderate-to-slow rate of speech of this study, where target undershoot is less likely to occur than in fast speech.

The analysis we conduct below examines the correlations between F0 peaks and valleys of each NP. Specifically, if correlation coefficients are relatively close to 0, it suggests that the F0 peaks and valleys are controlled independently, which follows more directly from the *target*-control hypothesis. In contrast, if the correlation coefficients are relatively far from 0, it supports the *register*-control hypothesis, because variation in register will result in non-independent (i.e. correlated) changes of peaks and valleys.

Furthermore, if we observe relatively strong peak-valley correlations, it is important to examine the signs of those correlations, because these allow for further inferences on which register parameter(s) – ceiling, floor, span – are the ones that are manipulated. Our interpretation of the correlation sign is premised on the assumption that the H and L targets are represented as fixed proportions of the register (F0 space) as in Figure 1, and that they are near the (normalized) ceiling and floor of the register. Under these assumptions, a *negative* correlation (the peaks and valleys tend to vary in the opposite directions) suggests that the speakers vary the register span to a greater extent than the register ceiling or floor. On the contrary, a *positive* correlation suggests that the speakers vary floor or ceiling more than span. It should be noted that there are more complicated ways for combinations of both span and floor/ceiling changes to result in positive or negative correlations, and our interpretations are limited to the simplest possibilities.

2. METHODS

2.1. Experiment design

A production experiment was conducted which elicited sentences of varying lengths. Specifically, the length of the subject phrase was varied so that it was composed of one, two, or three conjoined noun phrases (NPs). Each NP was comprised of a numeral (*eight, nine*), color (*red, green, blue*), and animal (*llamas, rhinos, weasels*). In cases of multiple-NP subject phrases, animals were always unique, while numerals and colors could be repeated. An example for the three NP sentences was "*Nine green rhinos and eight red weasels and eight blue llamas live in the zoo.*" All NPs were cued with visual stimuli.

For sentences with two and three NPs, a condition was tested in which the stimuli cueing the non-initial NPs were presented immediately after participants initiated an utterance rather than at the beginning of the trial. In these conditions, participants had

to quickly incorporate the newly presented NPs into their ongoing utterance. A total of five experimental conditions – 3DS (three subject NPs with delayed stimuli), 3NS (three NPs without delayed stimuli), 2DS, 2NS, and 1NS – were thus tested in the experiment, and the conditions were randomized from trial to trial. Each of the 3DS, 3NS, 2DS, and 2NS conditions appeared 45 times, and the 1NS condition appeared 90 times in each experimental session, which resulted in a total of 270 trials. 13 native speakers of English participated in the experiment. Note that the distinction between DS vs. NS conditions is indeed not necessarily relevant to the control hypotheses that we examine in this study; they exist to test hypotheses that are not considered here.

2.2. Data and measurements

Data were collected at a sampling rate of 22050 Hz. Acoustic segmentation was conducted using the Kaldi speech recognition toolkit ([12]). For each participant, ten trials, which included at least one instance of each numeral, color, and animal, were manually labelled to train monophone HMMs, and the rest of the trials were forced aligned.

Before conducting analyses, trials with potential disfluencies or with problems in data collection were removed. Trials with potential disfluencies were first algorithmically identified. Specifically, for each word and between-word silence interval, a mixed-effects linear regression model was fit to the word or interval durations with experiment condition as a fixed effect and participant as a random intercept. Trials with extreme durations were considered to contain disfluencies and excluded from subsequent analyses (265/2970 trials: 8.9%). Data from two participants were excluded due to a high proportion of hesitation or speech errors (i.e. more than 20% of trials). A small number of trials with problems in recording or stimuli presentation (i.e. delayed stimuli presented before utterance initiation) were also identified and discarded (12/2970 trials: 0.4%). This altogether left 2693 trials in total.

For the remaining trials, F0 data were extracted in Praat. A smoothed and interpolated F0 contour was then generated for each trial. To ensure that the accentual patterns in the analysis were comparable across participants, an average time-warped F0 contour was generated for each participant and condition and was qualitatively compared. Among 11 participants, seven of them showed similar intonation patterns in the subject phrase; specifically, they produced an F0 valley, F0 peak, and another F0 valley at each NP (which presumably constitutes a phonological phrase). F0 values of these landmarks of the seven participants were measured and subject to analyses.

2.3. Data analysis

The linear correlations between F0 peaks and valleys were calculated for each participant and NP. In each of these cases, we analyzed correlation between (i) preceding valley and peak, and correlation between (ii) peak and following valley. These correlations were calculated separately in DS and NS conditions. This is because variations in peaks and valleys may differ by whether all stimuli were presented at the beginning of the trial (fully planned before utterance initiation) vs. when some of them were delayed (had to change utterance plan after production). Note, however, that how the F0 values of peaks and valleys differ between DS vs. NS conditions is not directly relevant to the goal of the current study and is not analyzed here.

Thus, we obtained four correlation coefficients for each combination of participant and NP – i.e. two types of correlations (preceding valley & peak, peak & following valley) were calculated for each of the DS and NS trials. There were seven participants, and the subject phrase had maximum of three NPs, which resulted in a total of 84 correlation coefficient values.

3. RESULTS

The analysis showed that the peaks and valleys are highly correlated, providing evidence for the *register*-control hypothesis. In the majority of cases (i.e. in all participants/NPs, between peaks and preceding/following valleys), the correlation coefficients were positive and far from 0. Figure 2 shows the coefficients calculated with preceding valley and peak (a)/(b), and peak and following valley (c)/(d) at NP1 and NP2. Except for a handful of cases (e.g. (a) PA01, (b) PA02, (d) PA06), correlation coefficients were positive and relatively far from zero.

In fact, out of 84 correlation coefficients we have obtained, only five of them were negative. In addition, the absolute coefficient values were above 0.2 in 67 cases, showing that most of them were far from 0.

4. DISCUSSION AND CONCLUSION

This study aimed to assess evidence for two ways in which the control of F0 can be accomplished – *target*-control vs. *register*-control. We examined the

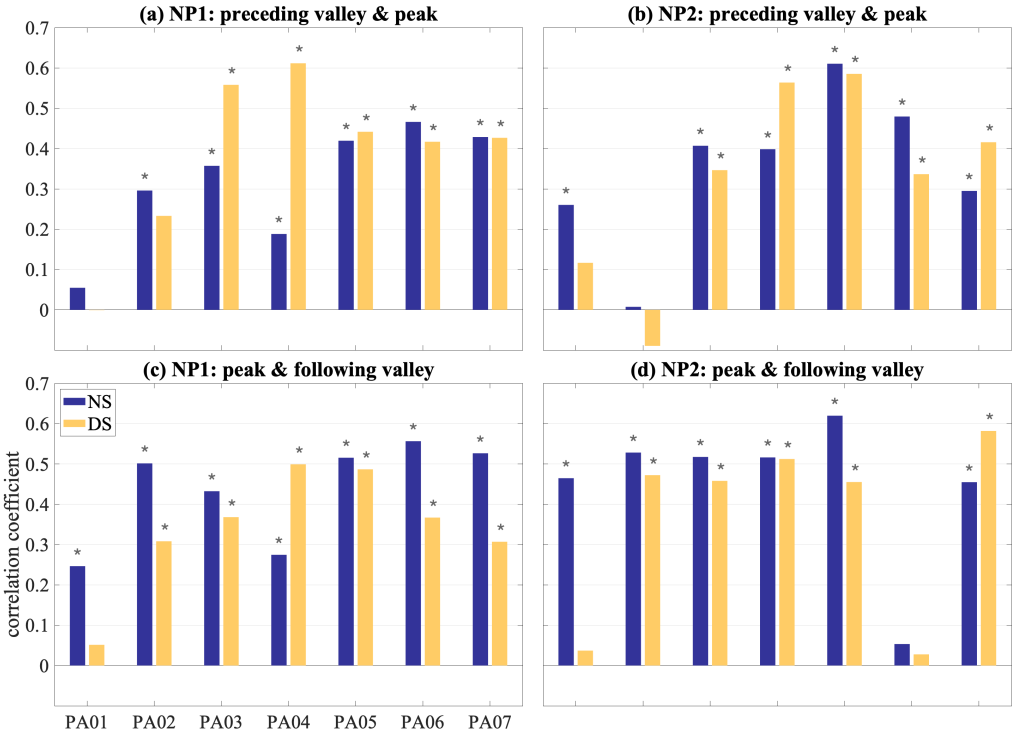

Figure 2: Correlation coefficients (a), (b): between preceding valley and peak, (c), (d): between peak and following valley at NP1 and NP2. An asterisk indicates a significant correlation ($p < 0.05$). The average number of data points used to calculate the correlation was 158 (NS) and 71 (DS) for NP1, and it was 73 (NS) and 70 (DS) for NP2.

correlations between F0 peaks and valleys at each NP, and the correlation coefficients in most cases were far from 0. This finding suggests that F0 peaks and valleys are not independently controlled but are correlated, providing evidence for the *register*-control hypothesis. A plausible interpretation of the correlation is that for a given utterance, speakers have a set of invariant cognitive representations of high and low pitch targets, and they control pitch register to realize the abstract representation into different F0 peaks and valleys (Figure 1-(ii)).

Furthermore, the analysis found that the F0 peaks and valleys are positively correlated in most cases (i.e. coefficients > 0). The simplest interpretation of this finding is that speakers manipulate either ceiling or floor to a greater extent than they manipulate span, at least in the current experimental task. The reasoning behind this interpretation is as follows: H and L pitch targets are likely to be located near the register ceiling and floor, respectively. If span is the primary register parameter that speakers manipulate, the contractions/expansions of span across NPs will lead to opposing changes in peaks and valleys, i.e. negative correlations. Conversely, if either the floor or ceiling is the primarily manipulated parameter, F0 peaks and valleys will tend to be mutually raised or lowered, i.e. positive correlations. We note that these inferences assume that *register*-control is as parsimonious as possible, in that speakers manipulate primarily just one of the three register parameters. Note that in the case of manipulating span, one or both of the edge parameters (ceiling and/or floor) would change indirectly. Conversely, in the case of manipulating the edge parameters, indirect span changes may occur.

Another caveat regarding the interpretation of our results is that there is a more complicated account of the empirical patterns that is consistent with *target*-control. Specifically, speakers might control H and L pitch targets in a correlated manner, not because of the register per se, but due to some other unknown mechanisms. Yet, we hold that the *register*-control interpretation is preferrable, since it does not require an additional mechanism for correlating H and L pitch targets. Moreover, there are a wide variety of phonological patterns (such as downstep and post-focus compression) which can be well understood as instances of *register*-control (e.g. [13], [14], [15]). Thus, although both *target*- and *register*-control can generate correlations of peaks and valleys, *register*-control is the simpler one and also has external motivations.

5. REFERENCES

[1] J. Pierrehumbert, "The phonology and phonetics of english intonation," Ph.D. dissertation, Massachusetts Institute of Technology, 1980.

[2] M. Liberman and J. Pierrehumbert, "Intonational invariance under changes in pitch range and length," in *Language sound structure*. MIT Press, 1984.

[3] J. Pierrehumbert and M. E. Beckman, *Japanese Tone Structure*. MIT Press, 1988.

[4] D. R. Ladd, *Intonational Phonology (Second Edition)*. Cambridge University Press, 2008.

[5] Y. Xu, "Speech melody as articulatorily implemented communicative functions," *Speech communication*, vol. 46, no. 3-4, pp. 220–251, 2005.

[6] G. Kochanski and C. Shih, "Prosody modeling with soft templates," *Speech Communication*, vol. 39, no. 3-4, pp. 311–352, 2003.

[7] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing," in *The Production of Speech*. Springer, 1983, pp. 39–55.

[8] ——, "Prosody, information, and modeling-with emphasis on tonal features of speech," in *Workshop on Spoken Language Processing*, 2003.

[9] D. R. Ladd, "An introduction to intonational phonology," in *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*. Cambridge University Press, 1992, pp. 321–334.

[10] G. N. Clements, "The description of terraced-level tone languages," *Language*, pp. 536–558, 1979.

[11] E. Gårding, "A generative model of intonation," in *Prosody: Models and measurements*. Springer, 1983, pp. 11–25.

[12] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[13] G. N. Clements, "The status of register in intonation theory," in *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*. Cambridge University Press, 1990, pp. 58–71.

[14] D. R. Ladd, "Phonological features of intonational peaks," *Language*, pp. 721–759, 1983.

[15] ——, "Metrical representation of pitch register," in *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*. Cambridge University Press, 1990, pp. 35–57.