# EVIDENCE FOR PREPLANNED AND ADAPTIVE F0 CONTROL

Seung-Eun Kim

Northwestern University
seungeun.kim@northwestern.edu

## ABSTRACT

This study investigates how speakers plan and adjust F0 in relation to the anticipated length of an utterance. Utterance length was manipulated prior to response initiation by cueing utterances with one, two, or three subject noun phrases (NPs) with visual stimuli. Furthermore, on some trials, a novel experimental manipulation was tested in which additional stimuli were presented upon detection of utterance initiation, thereby increasing the utterance length after participants began production. There are two main findings: (i) participants raised the F0 ceiling and floor and broadened the span of the initial NP when they had to produce longer utterances; (ii) participants reduced the amount of F0 ceiling compression from the first to the second NP to adapt to the changes in utterance length. Both findings suggest that speakers aim to maintain a sufficient F0 space while speaking, by taking into account the anticipated length of the utterance.

**Keywords:** F0 control, F0 preplanning, adaptive control, utterance length, delayed stimuli

## 1. INTRODUCTION

Many studies have investigated whether speakers vary utterance-initial F0 according to sentence length, aiming to find evidence for preplanning of F0 parameters. The hypothesis that was tested is that speakers would raise the initial F0 in longer utterances to avoid hitting the bottom of their F0 range before the utterance ends. A correlation between utterance length and initial F0 peak was examined in a variety of languages, and the studies have found mixed results. For instance, a significant effect of utterance length on the initial F0 peak was found in [1], [2], [3], [4], while no correlations were observed in [5], [6], [7], [8]. The results of the studies were inconsistent even within the same language (e.g. English: [1] vs. [5]).

The current study revisits the issue of F0 preplanning by examining the effects of utterance length on the overall F0 control, specifically by analyzing F0 peaks, valleys, and ranges. This contrasts with previous studies which almost exclusively examined F0 peaks. The goal of the current analysis therefore is to investigate how speakers control F0 register (F0 space) that is composed of F0 ceiling, floor, and span before utterance initiation (rather than just the ceiling), which may provide better insights on F0 preplanning.

The other question that is examined in this study is whether speakers can adjust F0 when utterance length changes after the start of production – i.e. adaptive F0 control. For this purpose, a novel experimental paradigm was developed in which the stimuli that cue the parts of the utterance are delayed until after participants initiate production. Under this condition, participants have to quickly adapt to changes in the length and content of the utterance.

As far as I am aware of, no studies have examined how speakers adapt F0 control to externally cued changes in the utterance length/content that occur after utterance initiation. These lexical/phrasal perturbations may be similar to or different from other sorts of perturbations made in the previous studies. For instance, [9] and [10] found evidence that speakers can incorporate a segment into an ongoing utterance that was missing before response onset but was cued after response initiation. A number of studies also found that speakers can adapt their F0 control to pitch-shifted auditory feedback (e.g. [11], [12], [13], [14], [15], [16]). In these studies, participants produced a vowel, syllable, or sentence, while hearing back their production with F0 perturbations – i.e. pitch of their voice was either raised or lowered compared to the original production. Participants were sensitive to the pitch-altered feedback and in general showed compensatory responses.

In sum, this study examines speakers' preplanned and adaptive F0 control through variations in utterance length. Participants produced sentences with one, two, or three subject noun phrases (NPs), which were cued by visual stimuli. For the multiple-NP sentences, two conditions were tested: one in which the stimuli for the non-initial NPs were delayed, and the other in which no stimuli were delayed. Various F0 parameters – peaks, valleys, and ranges – were examined for the length effects.

The following are the hypotheses and predictions.

**Hypothesis 1. Preplanned F0 control**: Speakers will make a pre-utterance F0 plan according to the initial utterance length. *Predictions*: The values of the F0 peaks and valleys will be higher, and ranges will be larger in longer utterances.

**Hypothesis 2. Adaptive F0 control**: Speakers will adapt F0 control in response to changes in utterance length. *Predictions*: The differences in the values of the F0 peaks, valleys, and ranges between the first and second NP will be smaller (less F0 register compression) in the condition where the stimuli are delayed compared to the condition without delay.

## 2. METHODS

### 2.1. Participants and experiment design

13 native speakers of English (7M, 6F) participated in the experiment. Participants produced sentences with one, two, or three subject NPs. The lexical and phonological content of the NPs was controlled as monosyllabic numeral {*eight, nine*} + monosyllabic color {*red, green, blue*} + disyllabic animal {*llamas, rhinos, weasels*}. In sentences with multiple NPs, animals were always unique, although numerals and colors could be repeated. Participants were instructed to connect NPs with the conjunction "*and*". An example sentence with three NPs was "*Nine green rhinos and eight red weasels and eight blue llamas live in the zoo*". In the experiment, NPs were cued with visual stimuli as in Figure 1.

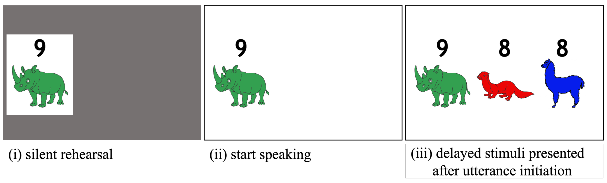

**Figure 1:** Presentation of a single trial.

For sentences with multiple NPs, a condition was tested in which the visual stimuli that cue non-initial phrases were delayed until after utterance initiation was detected. In this condition, participants saw only one stimulus before production, and the rest of the stimuli appeared immediately after they initiated an utterance. This condition is referred to as the *delayed stimuli* (DS) condition, as opposed to the *no-delayed stimuli* (NS) condition, in which all stimuli appeared before production. Thus, a total of five conditions were tested in the experiment – i.e. three lengths × two delayed conditions (1NS, 2NS, 2DS, 3NS, 3DS); cf. 1DS could not be tested.

In each block, there were 30 trials: each of the 2NS, 2DS, 3NS, and 3DS conditions appeared five times, and the 1NS condition appeared 10 times, all of which were randomized from trial to trial. Each experimental session had nine blocks.

The target sentences were cued as in Figure 1. In each trial, the initial stimuli appeared with a grey background (Figure 1-(i)). Participants were instructed to silently rehearse the sentence during this phase. The grey background lasted for 2.7s if there was a single stimulus, 4.4s for two stimuli, and 6.1s for three stimuli; these durations were determined based on the average utterance durations in the pilot experiments, which were conducted on three English speakers. After the given periods, the background automatically changed to white, which cued participants to start speaking (Figure 1-(ii)). In the DS conditions (2DS, 3DS), the stimuli that cued non-initial NPs were presented immediately after utterance initiation was detected (Figure 1-(iii)).

Participants were instructed not to emphasize any words in the utterance. This was to prevent them from putting contrastive focus on the parts of the NPs, which may affect natural intonation of the utterance. Participants performed 24 practice trials with balanced exposure of all conditions and words before the experiment session. This was to familiarize them with the task and to ensure that they produce sentences without any focus.

### 2.2. Data collection and processing

Acoustic signals were recorded at a sampling rate of 22050 Hz. Word and phone-level segmentations were conducted using Kaldi speech recognition toolkit ([17]). For each participant, ten trials, which included at least one instance of each numeral, color, and animal, were manually labelled and used to train monophone HMMs. A forced alignment was then conducted on the remaining trials. F0 data were extracted in Praat with participant-specific pitch floor and ceiling, and then a smoothed and interpolated F0 contour was generated for each trial. With the segmented data, F0 of the subject phrase was extracted for the analyses.

Due to the novelty of the delayed stimuli design, participants produced disfluencies such as hesitations or speech errors on some trials. These trials were identified algorithmically based on word and between-word interval durations. Specifically, for each word and interval, a mixed-effects linear model was fit to the durations with the experimental condition as fixed effect and the participant as random intercept. At each location, datapoints whose standardized absolute residuals were larger than 3.09 (the 0.1/99.9 percentiles of a normal distribution) were considered to be duration outliers, and trials with these outliers were excluded from the analyses. After excluding these trials, there were two participants for whom the exclusions constituted more than 20% of their data. This suggests that these participants did not conform to the task instructions, and thus, their data were excluded. In sum, among 2970 trials (270 trials × 11 participants), 265 trials (8.9%) were identified to contain duration-based outliers and excluded.

For those trials that had delayed stimuli, the timepoint of the presentation of the delayed stimuli and the onset of the utterance determined by the forced alignment were compared to ensure that the delayed stimuli appeared after utterance was initiated. Three trials (0.1%) were identified as errors, as the beginning of the utterance followed the presentation of the delayed stimuli. Nine trials (0.3%) that had problems in recording were also discarded. This left a total of 2693 trials (90.7%) for the analyses.

### 2.3. Data analysis

To ensure that the F0 patterns that are analyzed are similar across participants, a global analysis was first conducted to compare participant-level F0 patterns. An average time-warped F0 contour was obtained for each participant and condition and was qualitatively assessed. Out of 11 participants, seven of them had similar intonation patterns, in which an F0 valley occurred at the numeral, an F0 peak at the color, and another F0 valley at the second syllable of the animal. The current analyses thus focused on the F0 trajectories of these seven participants, and the F0 values of the landmarks – F0 valleys and peaks – and the ranges between them (F0 rises/falls) were measured.

For F0 preplanning, F0 values of the landmarks and rises/falls of the initial NP were analyzed using linear mixed-effects models. For each measurement, a mixed-effects model was fit to F0 values with the fixed effect of utterance length (i.e. number of subject NPs) and the random intercept of participants.

For adaptive control, stepwise regressions were conducted on the differences of F0 values of the landmarks and ranges between the first and the second NP (e.g. peak differences between NP1 and NP2). A mixed-effects model with the fixed effects of utterance length and delay, their interactions, and the random intercept of participants was fit to each of the difference measures. The maximal model was subsequently compared to the model that lacked the interaction term and the models that lacked either of the fixed effects. As delayed stimuli were presented shortly after response onset, their effects were likely to emerge during NP1 and NP2.

## 3. RESULTS

The results found evidence for both preplanned and adaptive F0 control. First, it was found that the participants started with a higher F0 peak and valley as well as a wider F0 range, when they were presented with more initial NPs, providing evidence for F0 preplanning. Not only the F0 peaks, but also valleys and ranges varied with utterance length.

For this analysis, the five experimental conditions were grouped according to the number of initial stimuli; thus, 1NS, 2DS, and 3DS were coded as 1Pi (one initial stimulus), and 2NS and 3NS were coded as 2Pi and 3Pi, respectively. This alternative coding is motivated by the finding that the 1Pi conditions (1NS, 2DS, 3DS) showed a similar F0 pattern in

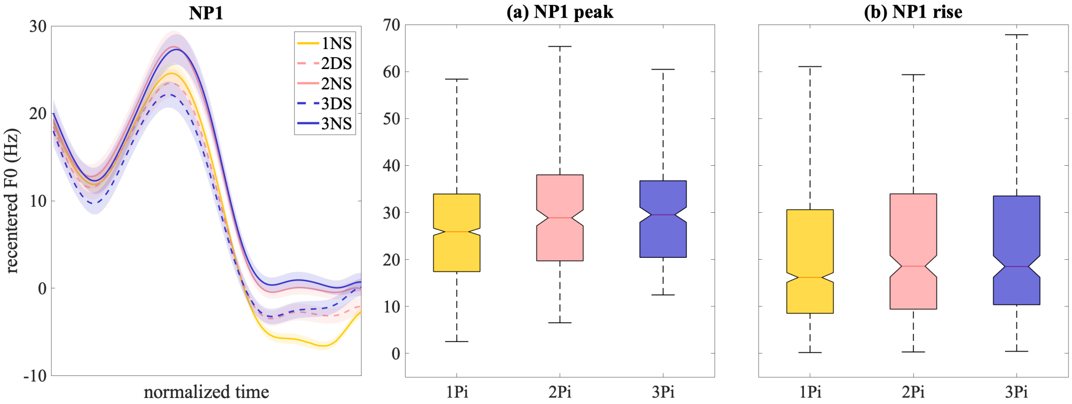

**Figure 2:** Average time-warped F0 contours of NP1 and the distributions of (a) peaks and (b) rises. For visualization, F0 (y-axis) was recentered within each participant using their global F0 mean.

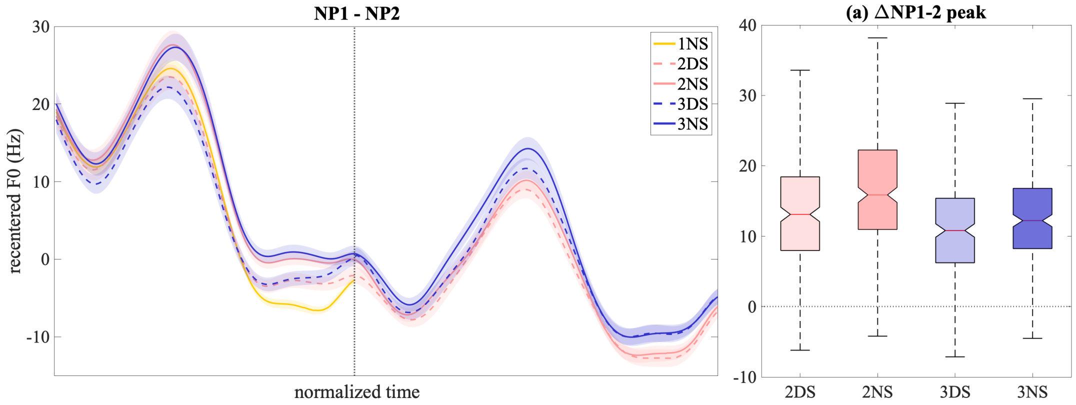

**Figure 3:** Average time-warped contours of NP1 and NP2 and the distributions of (a) peak differences. The vertical dotted line in the left panel represents NP boundary (i.e. the length of the time-warped NP).

NP1, which can be found in the leftmost panel of Figure 2. The regression coefficients of the 2Pi and 3Pi compared to 1Pi were 1.87 Hz and 2.07 Hz for F0 valleys ($p < 0.01$), 3.94 Hz and 5.06 Hz for F0 peaks ($p < 0.001$), and 2.11 Hz and 2.64 Hz for F0 rises ($p < 0.01$). Differences between conditions are also shown in Figure 2-(a)/(b). The difference between 2Pi and 3Pi was significant only at the peak and was minimal, as they differed only by 1.31 Hz ($p < 0.01$).

Second, participants lowered F0 peaks to a *lesser* extent when they saw delayed stimuli. Participants in general lowered F0 peaks from NP1 to NP2, as shown in the left panel of Figure 3. However, the amount of peak compression differed between DS and NS conditions such that the compression was larger in the NS conditions than the DS conditions. This is shown in Figure 3-(a) and confirmed by the statistical analysis, in which the regression coefficient of NS conditions (compared to DS) was 2.78 Hz for the F0 peak differences between NP1 and NP2 ($p < 0.001$). This suggests that as participants saw delayed stimuli, they adapted to the increase in the utterance length, by reducing the amount of F0 peak lowering (i.e. F0 register compression) that they would otherwise have done.

## 4. DISCUSSION AND CONCLUSION

This study examined whether speakers preplan F0 according to the initial utterance length, and moreover, whether they adapt F0 control in response to the unanticipated changes in length. The results found evidence for both preplanned and adaptive control: first, participants produced higher F0 peaks and valleys and wider F0 ranges when they had to produce longer utterances; second, when additional phrases were presented, they reduced the amount of F0 peak lowering to accommodate length changes.

One likely motivation behind both of these patterns is to establish a sufficient F0 space for production of upcoming phrases, given an expected declination of F0 across the utterance. When participants are producing a long utterance, they start from a higher F0 ceiling/floor as well as a wider F0 span, to avoid a circumstance in which the declination leads F0 floor to be too low or F0 span to be too narrow. When the length of the utterance changes after they start production, they quickly adjust the amount of F0 ceiling compression to make more F0 space for delayed phrases. The results thus suggest a strong tendency of speakers to produce F0 within their habitual, comfortable range and to establish sufficient F0 space throughout the utterance.

The present findings contribute to the inconsistent literature on F0 preplanning, by showing that the speakers factor utterance length into their F0 control prior to utterance initiation. It further informs us that it is not just the F0 peak, but other F0 variables such as valleys and ranges that are influenced by utterance length. Moreover, the adaptive responses suggest that speakers continuously estimate the remaining length of the utterance as well as the relation of their current F0 from the available F0 space and adjust their F0 control if necessary.

Overall, this study found robust evidence for F0 preplanning and adaptive control. Some possible future directions are whether similar results are observed when length is manipulated differently (e.g. adding sentences, words rather than NPs), whether there are individual differences in preplanned and adaptive F0 control, and whether the production differences observed in this study are perceptible by listeners.

## 5. REFERENCES

[1] W. E. Cooper and J. M. Sorensen, *Fundamental frequency in sentence production*. Springer Science & Business Media, 1981.

[2] G. Bruce, "Developing the swedish intonation model," *Working papers/Lund University, Department of Linguistics and Phonetics*, vol. 22, 1982.

[3] C. Shih, "A declination model of mandarin chinese," in *Intonation*. Springer, 2000, pp. 243–268.

[4] Y. O. Laniran and G. N. Clements, "Downstep and high raising: interacting factors in yoruba tone production," *Journal of phonetics*, vol. 31, no. 2, pp. 203–250, 2003.

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

[6] R. van den Berg, C. Gussenhoven, and T. Rietveld, "Downstep in dutch: Implications for a model," in *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*. Cambridge University Press, 1992, pp. 335–359.

[7] P. Prieto, C. Shih, and H. Nibert, "Pitch downtrend in spanish," *Journal of Phonetics*, vol. 24, no. 4, pp. 445–473, 1996.

[8] B. Connell, "Tone, utterance length and fo scaling," in *International symposium on tonal aspects of languages: With emphasis on tone languages*, 2004.

[9] D. H. Whalen, "Coarticulation is largely planned," *Journal of Phonetics*, vol. 18, no. 1, pp. 3–35, 1990.

[10] S. Tilsen, "Selection and coordination of articulatory gestures in temporally constrained production," *Journal of Phonetics*, vol. 44, pp. 26–46, 2014.

[11] T. A. Burnett, M. B. Freedland, C. R. Larson, and T. C. Hain, "Voice f0 responses to manipulations in pitch feedback," *The Journal of the Acoustical Society of America*, vol. 103, no. 6, pp. 3153–3161, 1998.

[12] J. A. Jones and K. G. Munhall, "Perceptual calibration of f0 production: Evidence from feedback perturbation," *The Journal of the Acoustical Society of America*, vol. 108, no. 3, pp. 1246–1251, 2000.

[13] T. M. Donath, U. Natke, and K. T. Kalveram, "Effects of frequency-shifted auditory feedback on voice f0 contours in syllables," *The Journal of the Acoustical Society of America*, vol. 111, no. 1, pp. 357–366, 2002.

[14] U. Natke and K. T. Kalveram, "Effects of frequency-shifted auditory feedback on fundamental frequency of long stressed and unstressed syllables," *Journal of Speech, Language, and Hearing Research*, vol. 44, pp. 577–584, 2001.

[15] S. H. Chen, H. Liu, Y. Xu, and C. R. Larson, "Voice f0 responses to pitch-shifted voice feedback during english speech," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 1157–1163, 2007.

[16] R. Patel, C. Niziolek, K. Reilly, and F. H. Guenther, "Prosodic adaptations to pitch perturbation in running speech," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 4, pp. 1051–1059, 2011.

[17] 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.