Imitation of intonational gestures: a preliminary report

Sam Tilsen, Danielle Burgess and Emma Lantz

1 Introduction

In this paper we report on an experimental investigation of how speakers imitate intonational gestures. The aim of the experiment was to determine which aspects of intonational contours are most directly controlled by speakers. Experiment participants heard parametrically varied HL (rising-falling) intonational contours in a synthesized trisyllabic name; they responded by imitating the name, embedding it in a fixed carrier phrase. Analyses of effects of stimulus parameters on responses show the following: (1) speakers control F0 targets of intonational tones, rather than the magnitudes or velocities of F0 rises and falls; (2) F0 targets are attracted to speaker-specific values, but can be modulated to achieve partial imitation; (3) speakers who imitate variation in the timing of pitch gestures do so with categorical changes in control over timing. These findings are argued to support a gestural model of F0 control, in which a HL tone is comprised of H and L tone gestures that are coordinated with one another and with other articulatory gestures.

1.1 Background

One of the main issues in modeling intonational F0 control has been whether the representations of intonational targets involve pitch changes or pitch levels. In other words, when speakers control F0 for intonational purposes, are they aiming to produce a pattern of F0 rises and falls, or are they aiming to achieve a sequence of F0 levels? The distinction has been a foundational issue in phonetic investigations of intonation, often referred to as "levels" vs. "configurations" viewpoints (Bolinger 1951). Ladd (2008) provides an extensive discussion of the historical and theoretical developments related to this distinction.

The configurations view holds that the targets speakers aim to achieve are patterns of rises and falls. In the early British tradition of intonational description (e.g. O'Connor, Arnold, & Arnold 1973), the emphasis was on a global pattern of changes in pitch, presupposing that patterns of F0 variation arise from continuous, undifferentiated control over F0. In the Dutch IPO (Institute for Perception Research) tradition (Cohen & Hart 1968), intonational patterns were associated with a discrete sequence of rises and falls. More recently, models of intonational control have been developed in which F0 contour shapes are generated by specifying a number of parameters, some of which influence rises and falls directly (Kochanski & Shih 2003; Xu & Wang 2001; Xu 1999). Configuration views predict that speakers should imitate stimulus parameters associated with rises and falls, such as the magnitudes or velocities of F0 changes.

In contrast, the levels view holds that the targets speakers aim to achieve are associated with F0 values. This notion is the basis for the autosegmental-metrical (AM) theory of intonation, which further holds that intonational contours arise from specification of a sequence of discrete, abstract H and L level tones. The AM theory in part originates from Bruce 1977, who showed that a precisely aligned peak is the most reliable correlate of word accent in Swedish. Bruce argued that tonal control required only specification of idealized F0 targets, with rises and falls being interpolations between them. Pierrehumbert (1980) developed a theory of English intonation along these same lines, and the levels view of the AM theory has taken hold in the dominant intonation transcription scheme, ToBI (Beckman, Hirschberg, & Shattuck-Hufnagel 2005; Silverman et al. 1992). The levels view predicts that speakers should imitate stimulus

parameters associated with F0 targets, rather than contour-based parameters such as the magnitude or velocities of F0 change.

Another important issue in modeling F0 control regards the timing of F0 patterns relative to segmental units or articulatory gestures. A number of studies have shown consistency in the alignment of F0 turning points to segmental boundaries (Arvaniti, Ladd, & Mennen 1998; Arvaniti & Ladd 1995; Prieto, Van Santen, & Hirschberg 1995), which is referred to as "segmental anchoring". Such patterns have moreover been found to be systematically influenced by phonological structure (Ladd, Mennen, & Schepman 2000; Prieto & Torreira 2007). Segmental anchoring implies that the targets of tones are anchored to some segmental boundary, such that the target is achieved at some point in time relative to (but not necessarily coincident with) the boundary (Ladd 2008). In order for a production mechanism to accomplish this sort of anchoring, explicit representations of when targets occur and when segmental boundaries occur are necessary.

An alternative perspective on F0 timing is based on the hypothesis that tones and intonational patterns are associated with pitch *gestures*, where such gestures are understood as analogous to oral articulatory gestures in the theory of articulatory phonology (Browman & Goldstein 1986, 1989). In articulatory phonology a gesture specifies a target value for a variable defined in coordinates of vocal tract geometry (such as lip aperture). When gestures are activated they drive movement toward that target, and words are associated with gestural scores that describe the timecourse of gestural activation. (Saltzman & Munhall 1989). Browman & Goldstein (1989) suggested that F0 changes could be associated with gestures, but this idea has not been investigated until recently. Gao (2008) demonstrated that tonal gestures in Mandarin are coordinated with consonantal and vocalic gestures. Two subsequent studies have shown that intonational gestures in German and Catalan are coordinated with vocalic gestures (Mucke, Nam, Hermes, & Goldstein 2012; Niemann, Mücke, Nam, Goldstein, & Grice 2011).

A gestural model of F0 control gels nicely with the AM theory of intonation, since H and L tones can be reconceptualized as gestural targets. One difference between pitch gestures and other articulatory gestures is that the targets of pitch gestures are most readily associated with an acoustic variable (i.e. F0), rather than a vocal tract geometry variable, as is the case for all other gestures. Whether or not this is problematic for attributing control of F0 to gestures is an open question. The discrepancy between how pitch gestures and oral articulatory gestures are conceptualized may be simply attributable to the practicality of measuring the articulatory variables involved in control of F0; alternatively, the task-dynamic model of articulatory phonology may need to be revised to allow for other sorts of targets. The gestural model entails that observations of segmental anchoring are indirect reflections of timing control: timing is accomplished through coordination of gestural initiation, and acoustic events like segment boundaries are generally not accurate indications of when articulatory movements are initiated. The gestural model predicts that imitations of alignment will reflect categorical changes in the structure of interaction between gestures, rather than gradient variation in alignment to segments.

There is a noteworthy conceptual disconnect between the gestural model of F0 control and the way in which segmental anchoring phenomena have generally been interpreted. Studies which have found evidence for the anchoring of an F0 turning point to some segmental boundary have generally interpreted this result as indicative of anchoring between the segmental boundary and an F0 *target*. Yet the standard articulatory phonology model of coordination incorporates no explicit mechanism for coordinating target achievement events with other gestural events. Turning points instead must be associated with points in time when gestures are initiated or deactivated, and the potential for gestural overlap obscures the relation between F0 turning points and target achievement (cf. also Silverman & Pierrehumbert 1990 for a discussion of this issue).

The problem arising from overlap is illustrated in Fig. 1. In a HL accent, by hypothesis comprised of a H gesture and a L gesture, the temporal location of the F0 peak is determined by when the L gesture is initiated. If the two gestures overlap to some extent, the H target may fail to

be achieved, and the F0 peak will occur earlier than it otherwise would. This phenomenon is likely related to tonal crowding, which has been observed when a nuclear accent is followed closely by another accent (Arvaniti, Ladd, & Mennen 2006; Silverman & Pierrehumbert 1990). In a monotonal H accent that is not subject to overlap or crowding, the gestural model holds that the location of the peak is determined by a return to neutral F0 value after deactivation of the H gesture. Only under the circumstance that the L gesture is initiated precisely when the H gesture achieves its actual target (dashed line in Fig. 1) does the turning point correspond to the H target achievement. Because gestural activation intervals are not directly observable, we cannot assess the extent to which gestures affecting F0 overlap. For this reason, the F0 turning point in a bitonal pitch accent cannot be assumed to index target achievement of the first tonal gesture; rather, the turning point indicates when the second gesture becomes active.

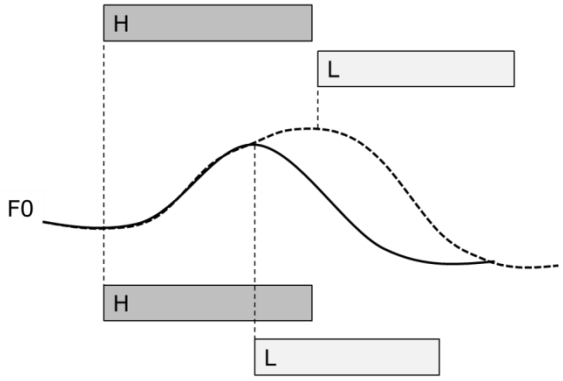

Fig. 1. Discrepancy between F0 turning points and tonal targets. Dashed line: with no overlap or underlap between H and L tone gestures the turning point corresponds to the onset of the L tone gesture. Solid line: with overlap between H and L tone gestures the turning point precedes the point in time when the H target would have been achieved.

The strategy adopted here to assess the predictions of a gestural model of F0 control involves probing how speakers imitate F0 contours. A number of previous experiments have used this approach. Pierrehumbert & Steele (1989) conducted an imitation task using synthetic stimuli in which the location of the F0 peak in the word *millionaire* (in the phrase "only a millionaire") was parametrically varied in 20 ms increments from 35 to 315 ms after the [m] (see Fig. 2). Durations of the rise and fall were held constant. Four of their five subjects exhibited bimodal distributions of peak locations. Pierrehumbert interpreted these patterns as indicative of a categorical distinction between L+H* and L*+H bitonal accent, associated with early and late peaks, respectively. Redi (2003) conducted a similar experiment, examining both F0 peaks and valleys. Her continua involved F0 extrema located in 25 ms steps across strong-weak and weak-strong sequences of syllables in synthetized words. The starting and ending F0 values flanking the target extremum were not varied, rather the velocities of the rise and fall were allowed to vary (see Fig. 2). Redi found a bimodal distribution of peak alignment, with the location of the category boundary differing between SW and WS disyllables. The results provide support for a categorical distinction between H+L* and H* accents. Dilley & Brown (2007) varied peak and valley timing by varying the relative levels of paired flat contours over a two-syllable sequence, and found evidence for a categorical shift from peaks timed on the first syllable to peaks timed on the second syllable. They also observed that speakers were accurate in imitating the F0 level ratios in a subset of their stimuli. Dilley (2010) used stimuli with a fixed, constant peak/valley and varied the F0 level preceding the rise/fall. Logarithmically transformed rise and fall intervals exhibited a highly linear relation to stimuli, rather than a bimodal distribution. Dilley argued that her results support the notion that rises and falls are comprised of two tonal targets, yet speakers have gradient control over pitch range.

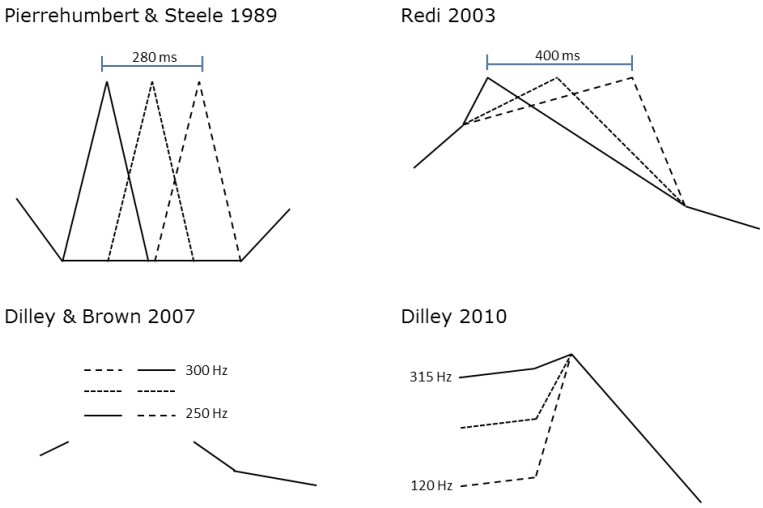

Fig. 2. Schematic representations of continua from pitch imitation studies. Only endpoints and midpoint of continua are shown. Manipulated portions of contours are contrasted with solid and dashed lines. Note that the continua in Dilley 2010 and Dilley & Brown 2007 involve logarithmically spaced frequencies, and that the range of the peak timing continuum in Redi 2003 differed across word shapes.

Despite the evidence for imitation of F0 parameters obtained in the aforementioned studies, none of these conclusively differentiates between levels and configurations models of F0 control. For example, in Pierrehumbert and Steele, the bimodal distribution of F0 peak timing could arise from a bimodal distribution of when the onset of the F0 rise occurs—because these two parameters covary the distinction regarding what speakers were controlling cannot be resolved. In Redi 2003, speakers may have imitated rise velocity and/or rise duration, both of which covaried with peak timing. In Dilley & Brown 2007 and Dilley 2010, speakers may have imitated rise and fall magnitudes and/or velocities, rather than levels. None of the reviewed studies dissociate configurational parameters from level parameters to the extent that effects of these stimulus parameters can be disentangled.

Other imitation studies that have been conducted show that speakers imitate F0 variation in stimuli and suggest that speakers maintain targets, or attractors, for intonational contours. Nonetheless, these studies do not directly resolve what the nature of those targets is. Braun, Kochanski, Grabe, & Rosner (2006) employed an iterative imitation paradigm in which speakers first imitate randomly generated F0 contours over utterances (based on three basis contours) and then iteratively imitate their imitations from the previous phase. They found that speakers gradually converge toward a set of distinct intonation patterns but retain some detailed variation through iterations that is not predictable from purely categorical phonological contrasts posited by the AM theory. Because the analysis is confined to global variables describing the contours it is unclear to what extent F0 parameters associated with more localized F0 changes were imitated. Nolan (2003) tested whether psychoacoustic pitch scales better account for imitation of pitch than Hertz; their subjects imitated naturally produced rising and falling contours produced in compressed, neutral, and expanded pitch ranges. They found that a semitone representation of rise/fall magnitudes resulted in the lowest observed error. Xu, Xu, & Sun (2004) used an "imitation via prosodic restoration" paradigm, in which speakers reproduce an intonational contour, portions of which they could not hear. They found that speakers produced raised and lowered F0 in association with focus contexts, despite not having heard the portions of the stimulus in which F0 was raised (on-focus) or lowered (post-focus).

1.2 Hypotheses

The current experiment required speakers to imitate a synthesized name in which a rise-fall F0 contour was parametrically varied. Speakers imitated the target word in a constant carrier phrase. The experiment varied the peak F0, timing of the peak, and starting or ending F0 in the stimuli. Hence in subsets of stimuli the range and velocity of the F0 fall are dissociated from peak F0, or the range and velocity of the F0 rise are dissociated from the peak F0. These dissociations and the variation in peak timing allow several hypotheses regarding the control of F0 to be tested.

F0 production models based on configurations hold that the targets of intonational gestures are the rises and falls of F0 contours, and so the control parameters in such models are the magnitudes and/or speeds of those rises and falls. Configurations models predict that speakers should imitate the magnitudes and speeds of rises and falls in stimuli rather than the levels of F0 peaks and valleys. In contrast, F0 production models based on levels hold that the targets of intonational gestures are F0 values (often simply a H and L tone). Such models predict that speakers should imitate the levels of F0 peaks and valleys rather than the magnitudes or speeds of rises and falls. Hence the two models provide competing hypotheses with conflicting predictions:

> HYP. 1A *Configurational targets:* speakers control the magnitudes and/or speeds of rises and falls in producing intonational gestures. PREDICTION: stimulus rise/fall range and/or speed parameters should account for more variance in response F0 characteristics than stimulus F0 peak/valley parameters.
>
> HYP. 1B *Level targets*: speakers control the levels of intonational gestures. PREDICTION: stimulus F0 peak/valley parameters should account for more variance in response F0 characteristics than stimulus rise/fall range and/or speed parameters.

A relevant questions that arises in either model is how flexible speakers are in adapting targets in an imitation task. One possibility is that speakers can arbitrarily alter targets; this predicts that the magnitude of variation in their imitations will match the magnitude of variation in stimuli parameters. A second possibility is that speakers have preferred targets—i.e. attractors in a control space; this predicts systematic deviations in responses toward a preferred target value, along with reduced magnitude of response variation relative to stimulus variation. Note that neither hypothesis predicts absolute imitation, since both allow room for perceptual reinterpretation of stimulus parameters or mapping to a speaker-specific control range.

> HYP. 2A *Flexible target representation:* speakers represent F0 targets in a flexible, gradient manner. PREDICTION: the magnitude of variation in response characteristics should match the magnitude of variation of? F0 in stimuli.
>
> HYP. 2B *Biased target representation:* the F0 target control space is biased toward preferred values but imitative mechanisms allow for modulation of those targets. PREDICTION: response F0 values should tend toward speaker-specific values, yet may exhibit partial dependence on stimulus parameters.

Recent experimental work has indicated that the onsets of intonational gestures (loosely, F0 turning points) are coordinated with the onsets of segmental articulatory gestures or the achievements of articulatory targets. To some extent acoustic segmental boundaries can serve as proxies for such articulatory events. If the coordinative relations between intonational gestures and articulatory gestures are invariant, the relative timing of F0 gestures and acoustic segmental

landmarks should be relatively constant across variation in stimulus peak timing or should reflect shifts among a set of categorically distinct coordinative patterns. Alternatively, if the timing of intonational gestures is not controlled through coordinative interactions between oral articulatory gestures and intonational gestures, then speakers should be able to imitate variations in stimulus peak timing linearly.

HYP. 3A *Gradient control over timing*: speakers control the timing of F0 peaks and valleys through some mechanism that allows for gradient specification of timing. PREDICTION: linear changes in stimulus peak timing will result in linear changes in response peak timing.

HYP. 3B. *Coordinative control over timing*: speakers coordinate the initiation of intonational gestures relative to oral articulatory gestures. PREDICTION: speakers will conform to one of two patterns: either there will be a single segment-tone onset interval which is least variable across all stimuli conditions, or speakers will exhibit a multimodal distribution of segment-tone onset relative timing.

2 Method

2.1 Stimuli

Stimuli were constructed using the Mbrola speech synthesizer (Dutoit et al. 1996; Dutoit, 1997). Mbrola is a diphone synthesizer that allows for parametric specification of segmental duration and F0 via PSOLA. American English voices *us2* and *us1* were used to synthesize male and female stimuli, respectively. All stimuli consisted of the segmental sequence shown in Table 1 below. Duration parameters were drawn from averages over several test productions of the target word produced in a carrier phrase with a H*L prenuclear accent by a trained phonetician, which results in a rising-falling contour. The segmental sequence was synthesized with 90 F0 contours by varying three of four parameters (cf. Table 2): the starting F0 value (F0 onset), the F0 value the peak (F0 peak), the timing of the peak relative to the onset of the target word (F0 peakt), and the ending value of the F0 contour in the word (F0 offset). The contours were generated by fitting smoothing splines to the points shown in Fig. 3. Participants were assigned to one of two groups, in which either the starting F0 or ending F0 was constant across all stimuli. Test productions of the target word were used to select a range of natural F0 parameters for a male voice, and parameters for the female voice were obtained by shifting all male parameters up 110 Hz.

Table 1. Segmental durations of the target stimulus

[m]	[a]	[n]	[i]	[m]	[ə]	
50	150	50	60	70	150	530 ms

Table 2. F0 contour parameters

		F0 onset	F0 peak	F0 peakt	F0 offset
Group 1	M	*110*	130, 140, 150	100, 150,	90, 100, 110
	F	*220*	240, 250, 260	200, 250,	200, 210, 220
Group 2	M	100, 110, 120	130, 140, 150	300 ms	*100*
	F	210, 220, 230	240, 250, 260		*210*

(italicized values were not varied in a given experiment)

Group 1: male stimuli Group 2: male stimuli

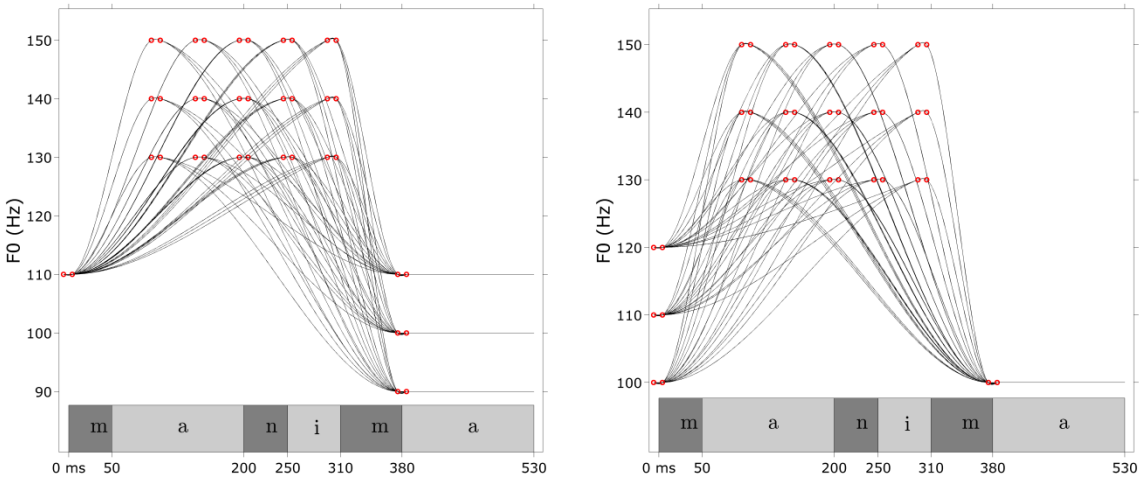

Fig. 3. Experimental stimuli F0 contours for male speakers. Group 1: F0 onset is fixed, F0 offset varies; group 2: F0 onset varies, F0 offset is fixed. There are 45 stimuli in each group.

2.2 Procedure

40 native speakers of English participated in a one-hour session (16 male, 24 female). Male and female participants imitated male and female voice stimuli, respectively. Half of the participants received stimuli with constant F0 starting level (group 1), the other half received stimuli with constant F0 ending level (group 2). Participants were seated in front of a computer monitor in a sound-attenuating booth and were recorded with a head-mounted microphone. Stimuli were delivered over computer speakers. The participants were instructed that they would hear the name "Manima" and then produce it in the sentence "we will lay Manima near a wall". They were told that they had two goals: to imitate the pitch of the word as accurately as possible, and to produce the sentence without hesitating before or after the target word. These two goals are to some extent in conflict, because producing the sentence with prosodic breaks before or after the target word can facilitate the task of imitating the target. To mitigate against this, participants practiced several trials of the experiment under experimenter supervision. If during the practice trials the experimenter judged the subject was hesitating within the sentence, they demonstrated how to produce the sentence without prosodic breaks and the subject practiced again until they were able to do so.

On each experimental trial, the speaker hears one of the stimuli and then produces the sentence, aiming to imitate the pitch of the stimulus. The 45 stimuli were presented in random order, distributed across successive blocks of 22 and 23 trials. After each trial (except the first five trials of the experiment) the speaker receives feedback regarding the accuracy of their imitation. The accuracy feedback was presented in the form of a vertical bar on the screen, whose height and color reflect an accuracy score. The score for each trial was calculated as follows. The raw F0 contour was extracted using the fxrapt function from the Voicebox toolbox (Brookes, 1997), which is based on normalized cross-correlation pitch tracking (Talkin, 1995). This contour was subsequently smoothed and extreme values were removed. The absolute difference between the stimulus contour on the target word and the sentence pitch contour was calculated from the beginning of the utterance at lags of 5 ms. Hence a minimum in this absolute difference function occurs at the lag where the response contour best aligns with the stimulus. The value of the difference function at this best-aligned lag was then taken as a raw imitation score. Then, the raw scores from all preceding trials were normalized through a z-transform, and the normalized score of the last trial was linearly mapped onto a score in the range of 1–100 by limiting its range to [-2,

2]. This scoring system ensures that scores are always relative to the speaker's own performance, and makes it more difficult to achieve high scores as accuracy improves. In addition, when a prosodic break was detected via the absence of voicing for more than 40 ms flanking the target word, speakers were warned to produce the target word without hesitating in the phrase—this served to further mitigate against the hesitation phenomenon described above. At the end of each block of trials, the speaker was presented with a numerical score, which was the average of their normalized scores in the last block of trials. Halfway through the session, participants were given a brief break. Most participants performed a total of 18–20 blocks, which amounts to 9 or 10 imitations of each unique stimulus.

2.3 Data analysis

Three of the 40 participants were excluded from analyses for failing to follow instructions or other reasons (i.e. falling asleep during the experiment, speaking with non-native accents, taking excessively frequent bathroom breaks). Data from four of the remaining 37 participants were excluded because the majority of their F0 contours lacked either the rise or fall component of the accent during the target word, and hence the F0 peak is not well-defined. Three of these produced F0 contours with only a rise in the target word, evidently delaying the fall until the end of the carrier phrase; these speakers may have interpreted the stimulus as a H* accent. One produced F0 contours beginning with a high F0 and exhibiting only a fall in the target word; this may indicate the speaker perceived the stimulus as a L* tone.

F0 contours were extracted from each response as follows. The recorded audio was high-pass filtered with a 3rd order elliptical filter having 70 and 125 Hz cutoffs for male and female speakers, respectively. The fxrapt function from the Voicebox toolbox (cite) was used to extract a raw F0 contour. For male speakers the allowable F0 range was 75–250 Hz, and for female speakers it was 125–450 Hz. The time frame was 11 ms and 8 ms for male and female speakers, respectively. Each contour was further processed by removing any F0 values exceeding 4 s.d. from the mean, interpolating any gaps (up to a maximum gap of 40 ms), and then fitting a smoothing spline. Occasionally creaky voice in responses prevented the pitch tracker from obtaining a reliable estimate of F0, hence trials with missing frames in the target word or preceding vowel were excluded (overall 2.2% of the responses).

Segmentation was conducted through forced-alignment using the HTK-HMM toolbox (Young et al., 2002). For each speaker, 10 randomly selected trials were hand-labeled in Praat. These were used to train HMMs and conduct a forced alignment on all trials. Subsequently HMMs were retrained on all responses in which no segmental duration exceeded 1.5 s.d. of the mean, and the data were re-aligned. The segmentation was then used to guide identification of tone gesture landmarks. These include the temporal locations and F0 values of the peak, preceding minimum, and following minimum. From these measures rise and fall durations, magnitudes, and average speeds (magnitude/duration) were calculated. The F0 onset, target, release, and offset landmarks were also identified, on the basis of 20% velocity thresholds in sigmoid functions fit to rising and falling portions of the contours. Independent variables in analyses are the stimulus onset F0, peak F0, peak timing relative to vowel onset, offset F0, F0 rise magnitude and velocity, and F0 fall magnitude and velocity. For each dependent variable, outliers ($> \pm 2.0$ s.d.) were excluded within subjects. For each combination of dependent variable and independent variable, a mixed effects linear regression with subject as a random factor was conducted. For the results reported in section 3.1, the marginal variance (Barton, 2013) was used to assess the proportion of variance explained by the independent variable (marginal R^2). Within subject regressions were also conducted for selected predictors.

3 Results

3.1 Effects of stimulus parameters on response F0 contours

Analysis of stimulus effects on responses supports the levels model (Hyp. 1B) over the configurations model (Hyp. 1A). Only two stimulus parameters, peak F0 and offset F0, had substantial effects on response F0 variables in the across-subjects analysis. Fig. 4 illustrates these effects, showing R^2 values > 0.10. Stimulus F0 peak accounted for 34% of the variance in response peak F0, 43% of the variance in the magnitude of the F0 rise, and 41% of the variance in F0 fall. Not surprisingly, F0 rise and fall velocities were influenced by stimulus peak F0 as well. However, stimulus peak F0 did not have an effect on onset F0 or offset F0. Stimulus offset F0 accounted for 43% of the variance in response offset F0, as well as a relatively minor portion of the variance in peak F0 and F0 fall. In all cases the directions of effects were consistent with imitation of the stimulus.

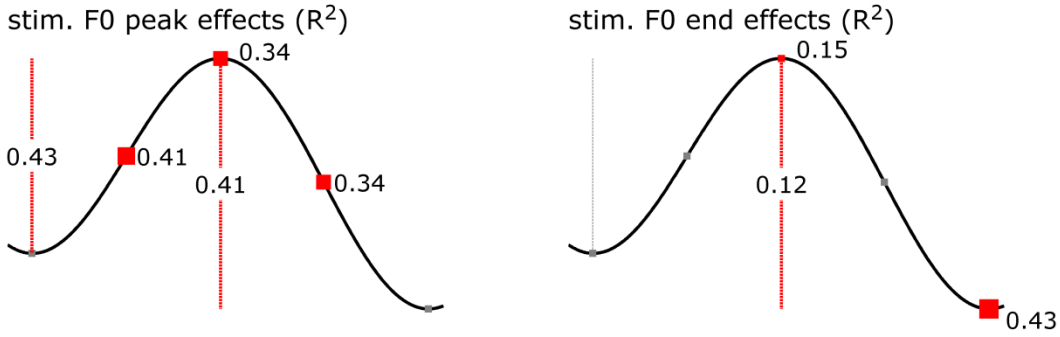

Fig. 4. Effects of stimulus F0 peak and F0 end on response F0 variables. Response variables for which predictors accounted for > 10% of variance are shown in red and labelled with the corresponding R^2. No other predictors beyond stimulus F0 peak and F0 end showed substantial effects across speakers.

None of the other stimulus parameters—i.e. onset F0, rise velocity, rise magnitude, fall velocity, fall magnitude, offset F0, and peak timing—accounted for more than a small portion of variance (10%) in response variables. The absence of substantial effects of configurational parameters (i.e. fall/rise velocities and magnitudes), in conjunction with the strongest effects being associated with F0 peak and valley parameters, supports a levels model over a configurations model. Speakers apparently adjusted their peak and offset F0 targets in attempting to imitate variation in the stimuli, and changes in response rise/fall characteristics were driven by those adjustments. The absence of effects of stimulus onset F0 is consistent with the assumption that speakers do not associate F0 prior to the accentual rise with an active tone gesture.

Despite the absence of substantial effects of stimulus peak timing in the across-subjects regression analysis, there were nonetheless some subjects who exhibited stronger effects of this parameter than others. The stimulus peak timing predictor had a significant effect on response peak timing (relative to vowel onset) for 22 of 39 subjects in the within-subject regressions. Of these subjects, 7 exhibited quite substantial effects where stimulus peak timing accounted for 40–70% of the variance in response peak timing, 5 subjects had moderate effects in the range of 10–25%, and 7 had relatively insubstantial effects. This suggests that some subjects were better able to imitate the timing of the peak than others. Imitation of peak-timing does not directly address the levels vs. configurations hypotheses, but does relate to results addressing coordination hypotheses that we consider further below.

3.2 Accuracy of imitation

Analyses of accuracy in imitation support a biased representation of onset and peak F0 targets (Hyp. 2B), rather than a flexible representation (Hyp. 2A). Fig. 5 shows mean response peak F0 for each of the three stimulus parameter values. In the left panel, speakers are sorted by the difference of their mean value from the central stimulus parameter. In the right panel, data were centered within speakers such that mean value of responses in the central stimulus parameter condition aligns with the central stimulus parameter, and speakers are sorted according to the sum of differences between stimulus values and mean response values for each stimulus parameter. Axis labels indicate F0 values for male and female speakers. The same centering and sorting schemes are employed in subsequent figures in this section. The raw peak F0 values are distributed above and below the range of stimulus parameters, suggesting that speaker-specific pre-existing F0 targets exerted a strong bias on F0 production targets. The centered data show that most speakers were able to partially adapt their F0 targets to achieve some degree of stimulus imitation. Several subjects exhibited a relatively high degree of target adjustment, but none of them exhibited the pattern of flexible adjustment predicted by Hyp. 2A. Most subjects exhibited partial adjustment that is more consistent with the predictions of Hyp. 2B, where subjects are strongly biased by a pre-existing production target. A similar pattern of partial adjustment is evident in F0 offsets, shown in Fig. 6. The re-centered values indicate that the majority of speakers produced some degree of F0 target imitation, yet were biased toward a speaker-specific value. For both F0 peaks and offsets, it is evident that speaker-specific variation contributes more than stimulus-induced variation.

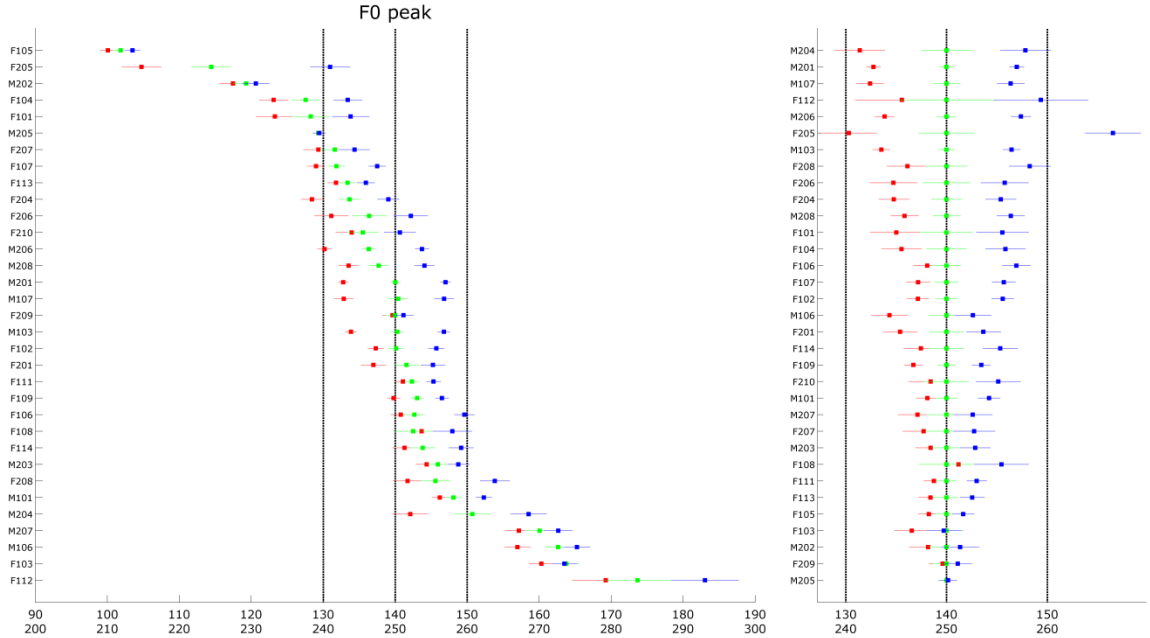

Fig. 5. Accuracy of response F0 peak by speaker. *Left*: raw F0 peak values for each stimulus peak F0, speakers sorted by proximity of mean value to the central stimulus value. *Right*: re-centered F0 values showing variation in the magnitude of imitation accuracy, speakers sorted by accuracy. Stimulus values are indicated by vertical lines.

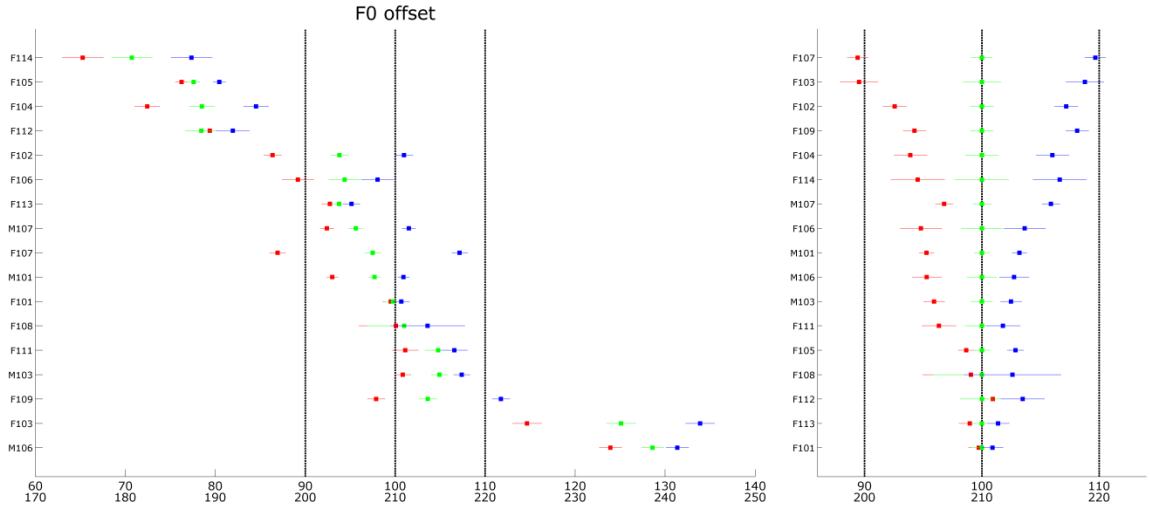

Fig. 6. Accuracy of response F0 offset by speaker. *Left*: raw F0 offset values for each stimulus F0 offset, speakers sorted by proximity of mean value to the central stimulus value. *Right*: re-centered F0 offset values showing variation in the magnitude of imitation accuracy, speakers sorted by accuracy. Stimulus values are indicated by vertical lines.

In contrast to peak and offset targets, F0 onset was not strongly imitated. The raw and re-centered onset F0 values in Fig. 7 show that very little imitation occurred, which is consistent with the regression findings reported above. The absence of strong imitation of F0 onset is expected on the basis of the notion that onset F0 is not associated with a tone gesture target. Instead, onset F0 may be attributable to a speaker-specific baseline F0.

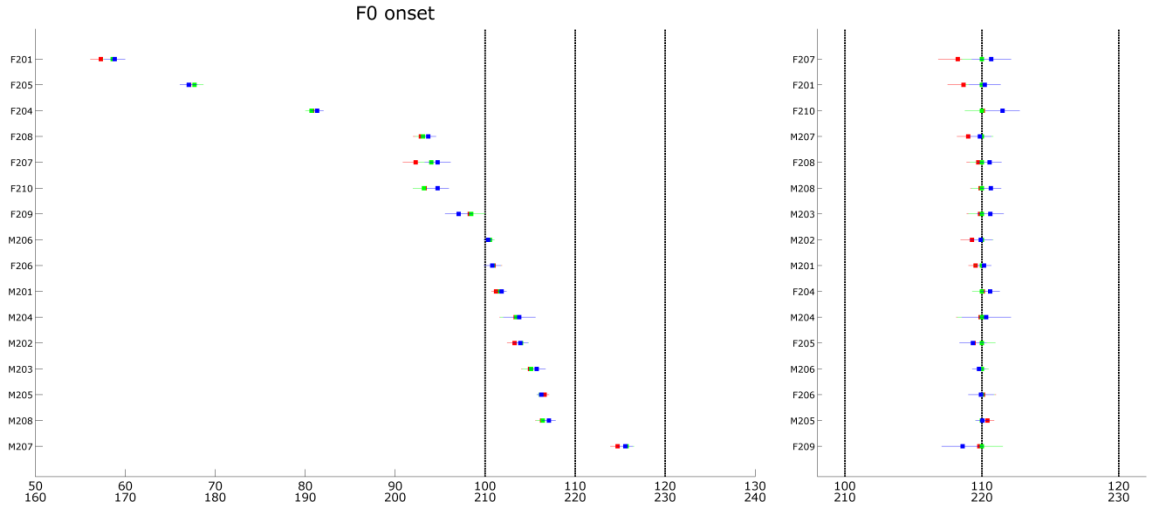

Fig. 7. Accuracy of response F0 onset by speaker. *Left*: raw F0 onset values for each stimulus onset F0, speakers sorted by proximity of mean value to the central stimulus value. *Right*: re-centered F0 onset values showing variation in the magnitude of imitation accuracy, speakers sorted by accuracy. Stimulus values are indicated by vertical lines.

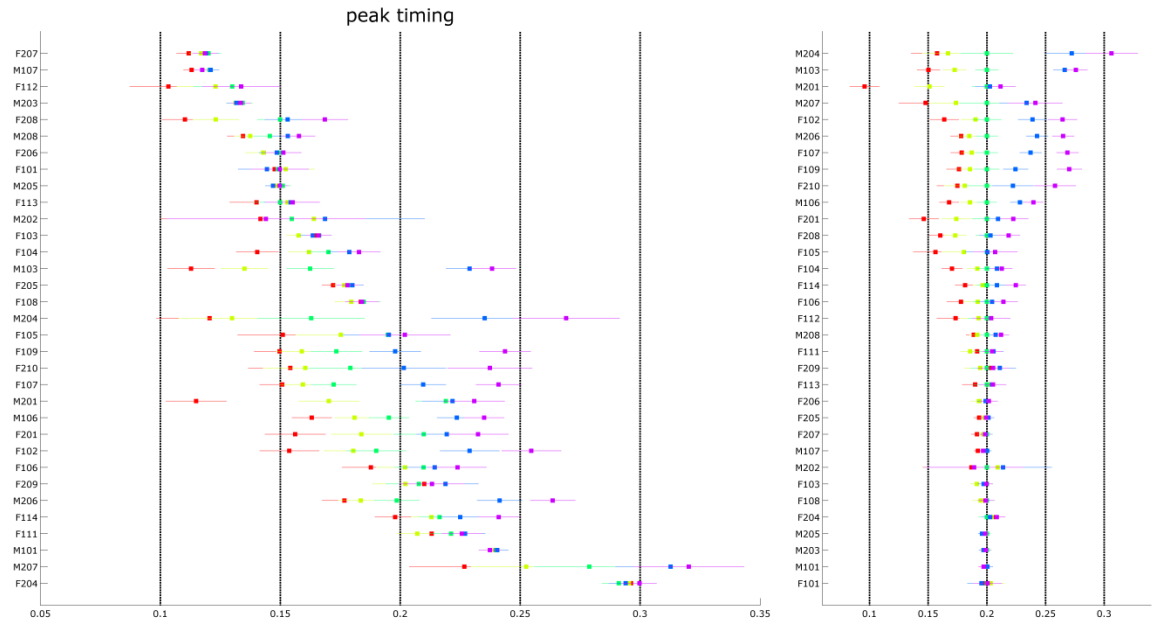

Fig. 8. Accuracy of response F0 peak timing by speaker. *Left*: raw F0 peak timing values (relative to target word onset) for each stimulus F0 peak timing, speakers sorted by proximity of mean value to the central stimulus value. *Right*: re-centered F0 peak timing values showing variation in the magnitude of imitation accuracy, speakers sorted by accuracy. Stimulus values are indicated by vertical lines.

The raw values of F0 peak timing in Fig. 8 show that the majority of speakers did not imitate this parameter. These speakers appear to be attracted to a value that may derive from coordinative relations between tonal gestures and articulatory gestures. However, re-centered data reveal that a minority of speakers exhibited partial imitation of peak timing. Moreover, a number of these imitating subjects show a pattern suggestive of bimodally distributed peak timing. We further explore these patterns below through analysis of timing between F0 extrema and segmental boundaries.

3.3 Timing of pitch gestures

Analyses of the timing of pitch gestures, relative to one another and to segmental boundaries in [manima], indicate that categorically different patterns of coordination were employed between and within speakers. Fig. 9 shows distributions of selected tone-segment lag distributions for each of the five stimulus peak timing conditions. In each panel the vertical line (time 0) corresponds to the start of the indicated segment. Speakers exhibited a strong tendency to align the onset of the H tone gesture to the word-initial [m]. The temporal location of the primary mode was unaffected by stimulus peak timing, occurring approximately 20 ms after the start of the [m]. This suggests that the H onset is coordinated with some gesture associated with the initial syllable (σ_1), although whether that gesture is the bilabial closure, the vocalic gesture, or the bilabial release cannot be determined. It is worth mentioning that the velocity-based threshold used here introduces some delay from the F0 turning point, so the true mode alignment may be up to 50 ms earlier. There also appears to be a secondary mode in the distribution later in the word. By examining the alignment of the H tone onset relative to the [n], this mode can be seen more clearly and appears to occur approximately 50–75 ms before the [n]. The secondary alignment mode is more pronounced in the latest stimulus peak timing condition, but nonetheless emerges in the other conditions. This suggests that H onset may have been coordinated with gestures in σ_2 by

some speakers, and that some speakers may have adopted this coordinative pattern in order to imitate stimulus peak timing.

Alignment of the L onset exhibited even more pronounced bimodality. The primary mode was fairly consistently aligned 10–20 ms before the [n] of the medial syllable (σ_2). In the first three stimulus peak conditions, the secondary modes are aligned just after [m] in the word-final syllable (σ_3). In the last two stimulus peak timing conditions, the primary modes are substantially lower in frequency and the corresponding secondary modes are aligned somewhat later, about 40–50 ms after the [m]. Taken together, these observations suggest that speakers vary regarding whether the L is coordinated with gestures associated with σ_2 or σ_3, and that some speakers may have shifted to coordinating the L gesture with σ_3 in order to imitate the stimuli with late peaks.

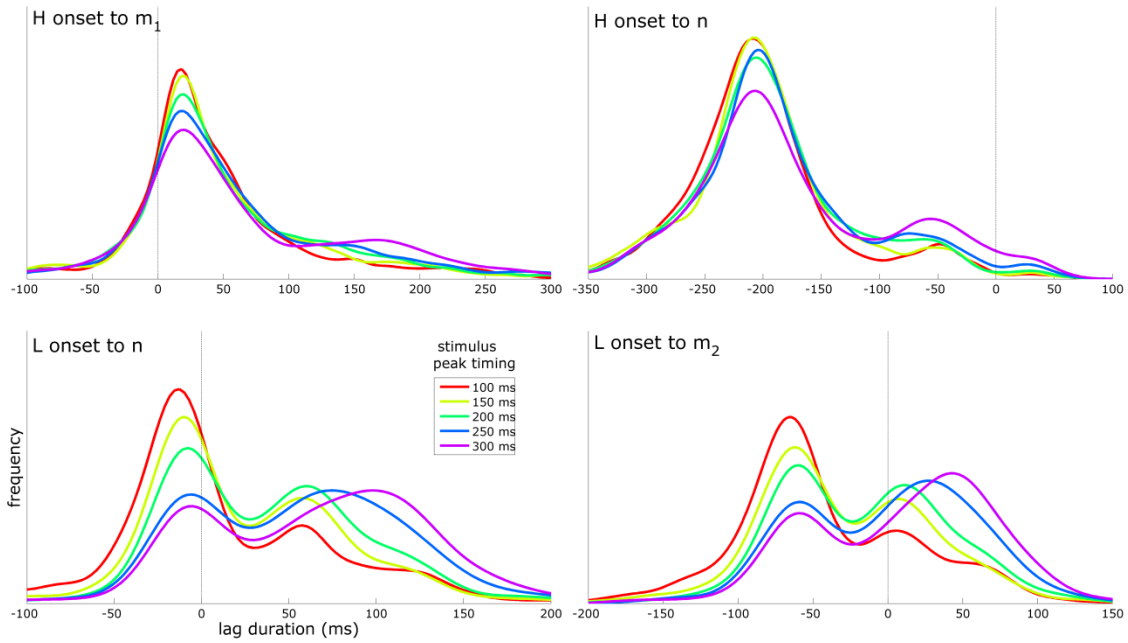

Fig. 9. Distributions of H and L tone onsets relative to selected segmental boundaries. Lines represent the frequency with which a particular tone-segment lag duration occurs across all responses in a given stimulus peak timing condition.

To further assess across-speaker variation in alignment and stimulus-induced variation, an analysis was conducted to identify segmental anchors, based on variability in intervals between H and L onsets and segmental boundaries. The analysis first identifies for each speaker and stimulus peak timing condition the least variable segment-tone interval for H and L tone onsets. For the L tone onsets the H to L onset interval is included as well. Then a regression is conducted between intervals associated with a common reference point, in this case the onset of [ei] in the carrier phrase. For example, when the H onset to [m] interval is the least variable for a given speaker in a given condition, then the [ei]-H onset interval is regressed by the [ei]-[m] onset interval. When the segmental interval accounted for more than 33% of the variance in the tone onset interval, the segment is considered to be an anchor for the tone interval, otherwise no anchoring is postulated. The anchoring can be viewed as indirect evidence for coordination. Tautosyllabic segmental anchors were combined in presenting the joint distribution of anchors, since onset consonantal and vocalic gestures are known to be coordinated and differences in variances associated with tautosyllabic tone-consonant and tone-vowel intervals were generally quite small.

Fig. 10 shows the joint distributions of anchors for the H and L gestural onsets, calculated separately for each stimulus peak condition. For the H tone onset, the majority of speakers

exhibited anchoring to σ_1 in all stimulus peak conditions (cf. the TOTAL column). There were a couple of instances in which σ_2 or σ_3 was an anchor for the H tone; these may reflect the secondary peak in Fig. 9. For about a third of the speakers there was no strong evidence for anchoring of the H tone in each stimulus peak condition. There appears to be a slight tendency for σ_2 or σ_3 to be more likely to be an anchor in the later stimulus peak timing conditions. These observations are consistent with the notion that a small minority of speakers coordinated the H tone with a non-initial syllable in all conditions, and some speakers switched to this coordinative mode when the stimulus peak occurred later.

For the L tone onset, it is evident that there was more across-speaker variation in anchoring. Overall the total numbers of minimally variable anchors associated with each syllable in the target word are comparable (cf. the TOTAL row in Fig. 10). Anchoring to σ_2 or σ_3 is expected on the basis of the two modes evident in Fig. 9: some speakers appear to coordinate the L with gestures in the medial syllable, others with gestures in the final syllable. However, the pattern in which the initial syllable anchors both the H and L tones is not entirely expected. In this pattern, despite that fact that the L tends to occur a couple hundred milliseconds later in the target word, the interval between the L and a σ_1 segment was less variable than any other interval. This pattern may be related to the one in which the H onset provides least variable anchor for the L onset. It suggests that for these speakers, the L gesture is coordinated with the H gesture, which may or may not in turn be anchored to gestures associated with σ_1. Furthermore, the tendency for stimulus peak timing condition to influence the anchor is stronger for the L tone than the H tone. The total distribution of L anchors can be seen to shift from σ_1 to σ_2 or σ_3 in the later peak-timing conditions.

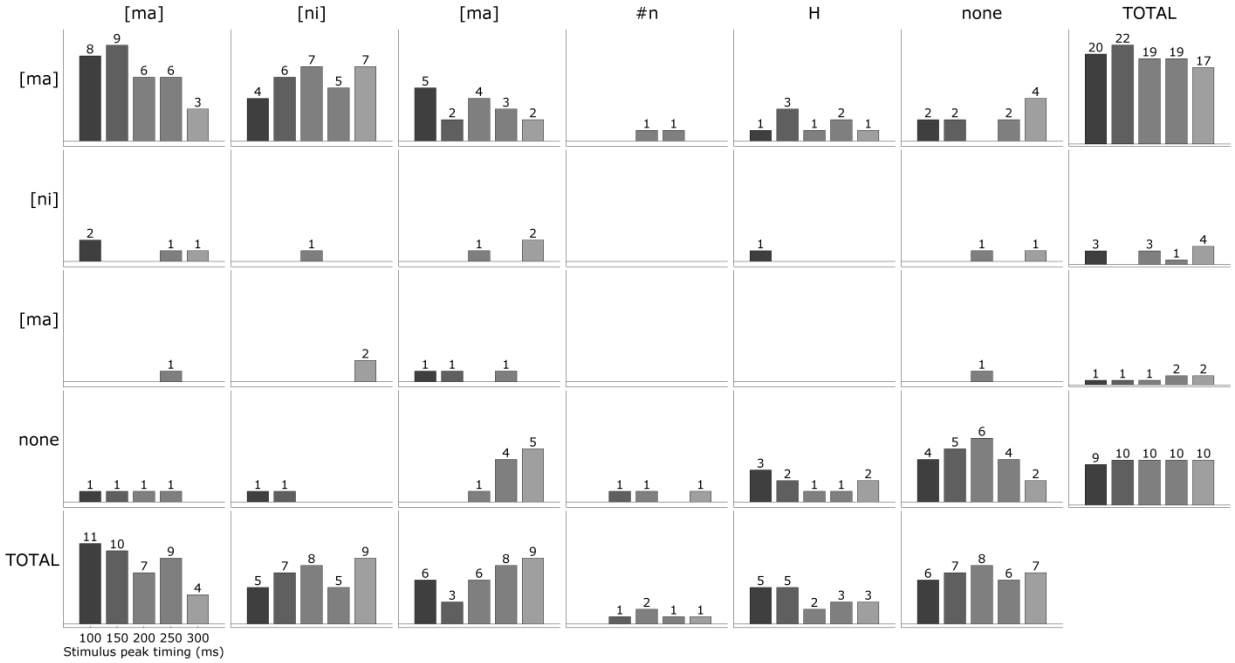

Fig. 10. Joint distribution of minimally variable anchors in each F0 peak timing stimulus condition. Rows represent anchoring for the H tone, columns represent anchoring for the L tone.

4 Discussion

The results of the current experiment indicate (1) that speakers control the F0 targets of tone gestures, rather than other parameters describing rises or falls; (2) that speakers are strongly biased toward individual-specific values for the targets of tone gestures; and (3) speakers can

adopt several modes of alignment/coordination of tonal gestures and segmental articulatory gestures. Below we discuss these findings further.

The analysis of effects of F0 parameters on response F0 variables in section 3.1 shows that only stimulus F0 targets had strong effects on response characteristics across speakers. This finding supports the notion that tones are gestures with F0 target parameters, which predicts that other sorts of parameters should not have a strong effect on imitation. The experimental dissociation of F0 rise/fall magnitude and velocity from F0 onset, peak, and offset values is essential to differentiating the configurations and levels hypotheses. As discussed above, previous designs have held one or both of these stimulus parameters fixed, and hence do not allow for their effects to be compared. Here we found that parameters associated with configurations do not substantially predict response characteristics, and hence the results favor the levels model over the configurations model. This leads to the inference that speakers have control over F0 targets, rather than contour parameters. It is also important to note that F0 onset, which is not assumed to be associated with an active gesture, did not strongly influence responses. Instead F0 onset may be associated with a baseline value that characterizes a speaker's range.

Several interesting questions regarding these results are deferred for future analyses or studies. For one, it would be desirable to test hypotheses regarding stimulus parameter interactions, i.e. whether models with combinations of F0 parameters outperform models with fewer parameters. Such analyses will be complicated by the number of interaction terms that must be considered and by the decision of whether to include speaker-specific random intercepts for interaction terms. Another avenue of future analysis involves transformations of response values. For example, the magnitudes of F0 rises and falls may be re-expressed logarithmically (cf. Dilley (2010); Nolan (2003)), and this may influence the pattern of observed results. Alternatively, the F0 contour across the entire carrier phrase, which tends to exhibit declination, may be used to normalize F0 within speakers so that rises and falls could be expressed relative to a dynamically varying baseline F0.

The second main finding is that despite imitating stimulus F0 targets to some degree, response F0 values were nonetheless strongly biased toward some speaker-specific value. This is evident in the figures in section 3.2, where re-centered values show that F0 peak and F0 offset were imitated to some degree, while raw values show that speaker-specific biases trumped the imitative adjustments. We speculate that this behavior results from integration of learned F0 targets associated with L and H gestures and a mechanism for task-specific modulation of targets. A simple way to conceptualize the integration is as follows.

Assume that the learned F0 target corresponds to a minimum of a quadratic potential function, as is the case for targets of vocal tract geometry associated with articulatory gestures. The task-specific modulation mechanism can then be modeled as the addition of a linear term with a slope parameter to the potential function, which will have the effect of shifting the minimum upward or downward, depending on the sign of the slope parameter. Exactly what the modulation corresponds to is an open question. Since F0 onset was not influenced by the stimuli targets, the modulation might be inferred not to reflect a baseline F0. However, if the modulation is active only when the tone gestures are active, then it would have no effect on F0 onset, which is consistent with the results. Alternatively, the modulation could reflect the expansion or contraction of pitch range, with the magnitude of the expansion or contraction related to the magnitude of the slope. This predicts that responses should be least variable where the stimuli parameters correspond to the average response value; subsequent analyses aim to test this prediction.

The third main finding of the current experiment is that there were categorical modes of tone-segment alignment, with speaker-specific and stimulus-induced variation in alignment modes. The H tone was most commonly aligned with the word-initial [m], although the distribution of lags in Fig. 9 shows that there was a minor secondary mode associated with alignment to the [n] of the word-medial syllable. Analysis of least variable anchors corroborates this observation, and

shows that stimulus peak timing had little effect on H alignment. In contrast, alignment modes for the L tone onset varied more across subjects: lag distributions show two comparable modes associated with [n] in the word-medial syllable and [m] in the word-final syllable [m]. Minimally variable anchor analyses show a more nuanced picture, where despite temporal proximity to segments in σ_2 or σ_3, a fair number of speakers appear to anchor the L tone to σ_1 or the preceding H tone. Furthermore, stimulus peak timing had a more pronounced effect on the distribution of anchors for the L tone than for the H tone: the distribution of L tone anchors shifts such that a greater proportion of anchors are associated with σ_2 or σ_3 in the later stimulus peak timing conditions.

The alignment patterns can be usefully interpreted in the context of a coordinative model of timing, where tone gestures are coordinated with oral articulatory gestures or other tone gestures. Because this experiment did not collect articulatory data, the results do not distinguish between a gestural model and a model in which tone targets are timed relative to segmental boundaries. Nonetheless, it should be noted that the identification of speakers who anchored the L tone to σ_1 or the H tone speaks to the gestural model. If the timing of F0 turning points is specified relative to segmental boundaries, then the least variable anchoring interval for a tone should correspond to some proximate boundary. Yet the H tone and σ_1 minimally variable anchors observed for? the L tone are relatively distal compared to other potential anchors. Hence the results are not consistent with a model which governs timing of tones through segmental anchoring; rather, they are amenable to understanding in a gestural framework where tones are coordinated with segmental gestures or other tone gestures.

References

Arvaniti, A., & Ladd, D. R. (1995). Tonal alignment and the representation of accentual targets. In *Proceedings of the XIIIth International Congress of Phonetic Sciences* (Vol. 4, pp. 220–223).

Arvaniti, A., Ladd, D. R., & Mennen, I. (1998). Stability of tonal alignment: the case of Greek prenuclear accents. *Journal of Phonetics*, *26*(1), 3–25.

Arvaniti, A., Ladd, D. R., & Mennen, I. (2006). Phonetic effects of focus and "tonal crowding" in intonation: Evidence from Greek polar questions. *Speech Communication*, *48*(6), 667–696.

Barton, K. (2013). Package "MuMIn."

Beckman, M., Hirschberg, J., & Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. *Prosodic Typology: The Phonology of Intonation and Phrasing*, 9–54.

Bolinger, D. L. (1951). Intonation: levels versus configurations. *Word*, *7*, 199–210.

Braun, B., Kochanski, G., Grabe, E., & Rosner, B. S. (2006). Evidence for attractors in English intonation. *The Journal of the Acoustical Society of America*, *119*, 4006.

Brookes, M. (1997). Voicebox: Speech processing toolbox for matlab. *Software, Available [Mar. 2011] from Www. Ee. Ic. Ac. Uk/hp/staff/dmb/voicebox/voicebox. Html*.

Browman, C., & Goldstein, L. (1986). Towards an articulatory phonology. *Phonology Yearbook*, *3*, 219–252.

Browman, C., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, *6*(2), 201–251.

Bruce, G. (1977). *Swedish word accents in sentence perspective* (Vol. 12). LiberLäromedel/Gleerup Malmo.

Cohen, A., & Hart, J. T. (1968). On the anatomy of intonation. *Lingua*, *19*(1), 177–192.

Dilley, L. (2010). Pitch range variation in English tonal contrasts: Continuous or categorical? *Phonetica*, *67*(1-2), 63–81.

Dilley, L., & Brown, M. (2007). Effects of relative F0 level on F0 extrema in an imitation task. *Ms. Bowling Green State University*.

Gao, M. (2008). *Tonal Alignment in Mandarin Chinese: An Articulatory Phonology Account*. Doctoral Dissertation, Yale University, New Haven, CT.

Kochanski, G., & Shih, C. (2003). Prosody modeling with soft templates. *Speech Communication*, *39*(3), 311–352.

Ladd, D. R. (2008). *Intonational phonology*. Cambridge University Press.

Ladd, D. R., Mennen, I., & Schepman, A. (2000). Phonological conditioning of peak alignment in rising pitch accents in Dutch. *The Journal of the Acoustical Society of America*, *107*, 2685.

Mucke, D., Nam, H., Hermes, A., & Goldstein, L. (2012). Coupling of tone and constriction gestures in pitch accents. *Consonant Clusters and Structural Complexity*, *26*, 205.

Niemann, H., Mücke, D., Nam, H., Goldstein, L., & Grice, M. (2011). Tones as Gestures: the Case of Italian and German. *Proceedings of ICPhS XVII*, 1486–1489.

Nolan, F. (2003). Intonational equivalence: an experimental evaluation of pitch scales. In *Proceedings of the 15th international congress of phonetic sciences* (Vol. 771, p. 774).
O'Connor, J. D., Arnold, G. F., & Arnold, G. F. (1973). *Intonation of colloquial English: a practical handbook* (Vol. 1). Longman.
Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation*. Massachusetts Institute of Technology.
Pierrehumbert, J., & Steele, S. A. (1989). Categories of tonal alignment in English. *Phonetica*, *46*(4), 181–196.
Prieto, P., & Torreira, F. (2007). The segmental anchoring hypothesis revisited: Syllable structure and speech rate effects on peak timing in Spanish. *Journal of Phonetics*, *35*(4), 473–500.
Prieto, P., Van Santen, J., & Hirschberg, J. (1995). Tonal alignment patterns in Spanish. *Journal of Phonetics*, *23*(4), 429–451.
Redi, L. (2003). Categorical effects in production of pitch contours in English. In *Proceedings of the 15th International Congress of the Phonetic Sciences* (pp. 2921–2924).
Saltzman, E., & Munhall, K. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, *1*(4), 333–382.
Silverman, K., Beckman, M. E., Pitrelli, J. F., Ostendorf, M., Wightman, C. W., Price, P., … Hirschberg, J. (1992). TOBI: a standard for labeling English prosody. In *ICSLP* (Vol. 2, pp. 867–870).
Silverman, K., & Pierrehumbert, J. (1990). The timing of prenuclear high accents in English. *Papers in Laboratory Phonology I*, 72–106.
Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). *Speech Coding and Synthesis*, *495*, 518.
Xu, Y. (1999). Effects of tone and focus on the formation and alignment of f0 contours. *Journal of Phonetics*, *27*(1), 55–105.
Xu, Y., & Wang, E. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication*, *33*(4), 319–337.
Xu, Y., Xu, C., & Sun, X. (2004). On the temporal domain of focus. In *Speech Prosody 2004, International Conference*.
Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., … Woodland, P. (2002). The HTK book. *Cambridge University Engineering Department*, *3*, 175.

Sam Tilsen
Department of Linguistics
Cornell University
203 Morrill Hall
159 Central Ave.
Ithaca, NY 14853
tilsen@cornell.edu