IMPLICIT IMITATION OF INTONATION CONTOURS IN WORD SHADOWING

Yu Jin Song, Cynthia G. Clopper

The Ohio State University
song.1693@osu.edu, clopper.1@osu.edu

ABSTRACT

Phonetic imitation of a range of segmental and suprasegmental features has been observed across tasks and instructions, suggesting that this process may be spontaneous and implicit. In contrast, imitation of phonetic features of intonation contours may depend largely on the explicitness of the instructions, especially in non-interactive tasks. In the current study, a word shadowing task without explicit instructions to imitate was used to explore the imitation of intonation contours and their phonetic implementation in American English. The analyses indicate imitation of rising contours, their f0 range, and their alignment with the stressed vowel. However, imitation of falling contours was limited to their alignment with the stressed vowel. These results suggest that the shape of intonation contours may affect their imitation, and that a perceptually more salient contour, such as a rise on an isolated word in American English, may lead to greater imitation.

Keywords: phonetic imitation, intonation, word shadowing, phonological abstraction

1. INTRODUCTION

Humans align their speech with the phonetic realization of the speech of their interlocutor [1, 2]. This phonetic imitation has been observed for a range of segmental features, including vowel quality [2, 3, 4] and voice onset time [5, 6], and for suprasegmental features, including overall f0 [1, 7, 8, 9] and speaking rate [7]. Talkers imitate in conversational settings [7, 10, 11] and in non-interactive tasks, such as word shadowing [1, 8, 12]. In non-interactive tasks, imitation has been observed both with and without explicit instructions to imitate [1, 6] and even with explicit instructions not to imitate [4], suggesting that phonetic imitation may be spontaneous and implicit.

In contrast, imitation of the phonological and phonetic features of intonation contours may depend on the explicitness of the instructions, especially in non-interactive tasks. In interactive tasks, talkers imitate both phonological intonation contours [10, 13] and their phonetic features, such as pitch scaling [14], without instructions to imitate. For example, Lee et al. [14] found imitation of rising boundary tones and of f0 maxima and minima in American English, using a maze navigation task.

In non-interactive tasks, talkers likewise imitate both phonological and phonetic features of intonation contours when they are explicitly asked to imitate [9, 15, 16, 17]. For example, Cole and Shattuck-Hufnagel [16] asked participants to repeat the way the sentence-length utterances were said and observed imitation of phonological features of the intonation contours, including pitch accents and boundary tones, but idiosyncratic imitation of phonetic features of the contours, such as the duration of pauses and glottalization at prosodic boundaries. Similarly, D'Imperio et al. [15] asked Italian participants to imitate as closely as possible the productions of a model talker who spoke an unfamiliar Italian dialect. Along with imitation of phonological features of the contours, D'Imperio et al. [15] found imitation of phonetic features of the contours, such as tonal alignment and pitch scaling. In non-interactive tasks without explicit instructions to imitate, talkers imitate intonation contours [18] and overall f0 [8, 9, cf. 19, 20], but implicit imitation of phonetic features of intonation contours has not been examined. Thus, convergence to phonetic features of intonation contours can be achieved by explicit instructions to imitate [15, 17]. However, imitation of phonological features of intonation contours tends to be more robust than that of phonetic features [16, 17, 21], suggesting that explicit imitation is mediated by abstract phonological representations [22].

In the current study, the goal was to further explore the implicit imitation of intonation contours and their phonetic implementation using a word shadowing task without explicit instructions to imitate. This approach allowed us to determine to what extent talkers spontaneously align their intonation contours with that of a model talker. We predicted that we would observe robust imitation of overall intonation contours [18], but less imitation of the phonetic features of those contours [16, 17, 21].

2. METHODS

2.1. Participants

Word shadowing data were collected from 45 participants (22 female, 23 male) recruited from a

local science museum in Columbus, OH. The participants were all native American English speakers and self-reported no history of speech, hearing, or language disorders. Their ages ranged from 19 to 69 years old (*M* = 35 years).

2.2. Stimulus materials

The stimulus materials comprised 48 multisyllabic English words, each produced by two female model talkers. The target words were 2-4 syllables long and the stressed syllable in each target word contained one of the following eight vowels: /i ɪ ɛ æ ɑ aɪ oʊ u/, with six words for each vowel. The auditory stimuli were selected from the Indiana Speech Project corpus [23]. One of the model talkers was from Indianapolis, IN, and was 19 years old at the time of recording. She produced all 48 target words with a rising intonation contour. The other model talker was from Fort Wayne, IN, and was 22 years old at the time of recording. She produced all 48 target words with a falling contour.

2.3. Procedure

Participants were seated at an individual testing station with a computer and a headset microphone in a glass-enclosed laboratory inside the science museum. In the first block of the word shadowing task, participants were asked to read aloud the set of words from the computer screen. Each of the 48 words appeared on the computer screen one by one. Each word stayed on the screen for 3.5 s followed by a .5 s blank screen to signal the onset of the next trial. Word order was randomized separately for each participant. This task served as the baseline for the shadowers' intonation contours.

A shadowing task followed in the second block. During shadowing, participants were asked to repeat the same set of words after one of the model talkers. Nineteen participants completed the shadowing task with the model talker who produced rising contours and 26 participants completed the shadowing task with the model talker who produced falling contours. Contour condition (Rise, Fall) was therefore a between-participant variable. In the shadowing block, the words were played one at a time over the headphones. Each auditory stimulus was preceded by a fixation cross which stayed on the screen for .5 s. The cross remained for another 4 s as the word played and the participants produced their repetition. A blank screen followed for 1 s to signal the onset of the next trial. Word order was randomized separately for each participant. Participant utterances were recorded in Audacity. All misread and misheard tokens were discarded prior to analysis. Thirty-nine tokens (2.14%) from the Rise condition and 43 tokens (1.72%) from the Fall condition were excluded.

2.4. Analysis

2.4.1. Contour coding

To assess imitation of the phonological features of the intonation contours, the intonation contours of the target words produced by the model talkers and by the shadowers in both the baseline and shadowing blocks were coded by a team of trained undergraduate research assistants. Each utterance was coded based on auditory and visual inspection as having been produced with one of three nuclear contour categories (rise, fall, plateau) to capture the intonation contour from the stressed syllable to the end of the word. The data from four shadowers (9%) were coded independently by two coders to assess reliability. The coders agreed on 97% of the utterances, suggesting very high reliability for this coding scheme.

Given that the model talkers produced exclusively rising contours in the Rise condition and falling contours in the Fall condition, respectively, imitation of intonation contours was assessed by comparing the proportion of rising contours in the baseline and shadowing blocks in the Rise condition and the proportion of falling contours in the baseline and shadowing blocks in the Fall condition. An increase in these proportions from baseline to shadowing indicates imitation (i.e., the proportions for the shadowers were closer to those for the model talker in the shadowing block than in the baseline block).

2.4.2. Acoustic analysis

To assess imitation of the phonetic features of the rising contours in the Rise condition and of the falling contours in the Fall condition, the f0 range and alignment relative to the stressed syllable of the rises produced in the Rise condition and of the falls produced in the Fall condition were analyzed. For all of the model talkers' utterances and for utterances produced by the shadowers in the baseline and shadowing blocks that were coded as rises in the Rise condition and as falls in the Fall condition, the recordings were analyzed in VoiceSauce [24] to extract f0 values at 1 ms intervals using the STRAIGHT algorithm [25].

The f0 range of the rises was defined as the difference between the minimum f0 value in the stressed vowel and the maximum f0 value between the minimum f0 location and the end of the word. The f0 range of the falls was defined as the difference between the maximum f0 value in the stressed vowel and the minimum f0 value between the maximum f0 location and the end of the word. The minimum and maximum f0 values were inspected for outliers. Values that fell outside of two standard deviations of an individual talker's mean were replaced with the

respective talker f0 maximum or minimum mean. In the Rise condition, 36 minima and 44 maxima (4.64%) were replaced and, in the Fall condition, 42 minima and 46 maxima (3.29%) were replaced. The corrected f0 maxima and minima were used to calculate f0 ranges in semitones to normalize for overall f0 range differences across talkers. Distances in f0 range between the shadowers and the model talkers were defined as the absolute difference in f0 range between the shadower and the model talker, separately for each word in each block [3, 19]. Given that f0 range measures were only available for rises produced in the Rise condition and for falls produced in the Fall condition, different numbers of tokens were included in each of the two blocks for each shadower. A difference-in-distance measure [3, 19] was therefore not possible. Imitation was instead assessed by comparing the f0 range distances in the baseline and shadowing blocks. A decrease in these distances from baseline to shadowing indicates imitation (i.e., the f0 ranges of the shadowers were closer to those of the model talker in the shadowing block than in the baseline block) [3, 19].

Alignment was defined as the time from the onset of the stressed vowel to the f0 minimum in the Rise condition and to the f0 maximum in the Fall condition. Similarly to the f0 range analysis, distances in f0 alignment between the shadowers and the model talkers were defined as the absolute difference in f0 alignment between the shadower and the model talker, separately for each word in each block [3, 19]. A decrease in these distances from baseline to shadowing indicates imitation [3, 19].

3. RESULTS

3.1. Rise condition

Table 1 shows the mean proportions of rising contours in the baseline and shadowing blocks in the Rise condition. To explore the effects of block (baseline vs. shadowing) and shadower gender (female vs. male) on the use of rising contours (rising vs. non-rising) in the Rise condition, a logistic mixed-effects model with the maximal random-effects structure that converged was constructed in R using lme4 [26]. Random effects were random intercepts by shadower and word and random slopes for condition by shadower and for gender by word. The main effect of block was significant ($\chi^2(1) = 10.44$, $p = .001$), confirming imitation of the model talker's rising contours in the shadowing block. The effect of gender and its interaction with block were not significant.

Table 2 shows the mean f0 range distance in semitones between the shadowers and the model talker in the baseline and shadowing blocks in the Rise condition. To explore the effects of block and shadower gender on the f0 range distances, a linear mixed-effects model with the maximal converging random-effects structure, which included random intercepts by shadower and word and a random slope for gender by word, was constructed in R using lmerTest [27]. The main effect of block was significant ($\chi^2(1) = 5.09$, $p = .024$), confirming imitation of the model talker's f0 range in rising contours in the shadowing block.

Baseline	Shadowing
.34 (.33)	.63 (.34)

Table 1: Mean (standard deviation) proportion of rising contours by block in the Rise condition.

Baseline	Shadowing
3.08 (1.19)	2.65 (.90)

Table 2: Mean (standard deviation) f0 range distance in semitones between the shadowers and the model talker by block in the Rise condition.

Table 3 shows the mean f0 alignment distance in milliseconds between the shadowers and the model talker in the baseline and shadowing blocks in the Rise condition. To explore the effects of block and shadower gender on the f0 alignment distances, a linear mixed-effects model with the maximal converging random-effects structure, which included random intercepts by shadower and word, was constructed. The main effect of block was marginally significant ($\chi^2(1) = 3.71$, $p = .054$), suggesting modest imitation of the model talker's f0 alignment in rising contours in the shadowing block.

Baseline	Shadowing
56 (15)	50 (13)

Table 3: Mean (standard deviation) f0 alignment distance in milliseconds between the shadowers and the model talker by block in the Rise condition.

3.2. Fall condition

The mean proportions of falling contours increased from 0.53 (*SD* = .39) in the baseline block to 0.55 (*SD* = .38) in the shadowing block in the Fall condition. To explore the effects of block and shadower gender on the use of falling contours (falling vs. non-falling), the same logistic mixed-effects model with the maximal converging random-effects structure, which included random intercepts by shadower and word and a random slope for block by shadower, was

constructed, as for the Rise condition. No effects were statistically significant.

The mean f0 range distance between the shadowers and the model talker increased from 2.85 semitones (*SD* = 2.23) in the baseline block to 2.97 semitones (*SD* = 1.04) in the shadowing block. To explore the effects of block and shadower gender on the f0 range distances, the same linear mixed-effects model with the maximal converging random-effects structure, which included random intercepts by shadower and word and a random slope for block by shadower, was constructed, as for the Rise condition. No effects were statistically significant.

Table 4 shows the mean f0 alignment distance in milliseconds between the shadowers and the model talker in the baseline and shadowing blocks in the Fall condition, separately for female and male shadowers. To explore the effects of block and shadower gender on the f0 alignment distances, a linear mixed-effects model with the maximal converging random-effects structure, which included random intercepts by shadower and word and a random slope for block by word, was constructed. The main effect of block was significant ($\chi^2(1) = 6.73$, $p = .009$), confirming imitation of the model talker's f0 alignment in falling contours in the shadowing block. The interaction between block and gender was also significant ($\chi^2(1) = 7.02$, $p = .008$). Post-hoc pairwise least square means comparisons revealed that the mean f0 alignment distance decreased significantly from the baseline block to the shadowing block for the female shadowers only ($t(208.2) = 3.76$, $p < .001$).

	Baseline	Shadowing
Female	49 (17)	36 (12)
Male	37 (14)	35 (32)

Table 4: Mean (standard deviation) f0 alignment distance in milliseconds between the shadowers and the model talker by block and gender in the Fall condition.

4. DISCUSSION

The results of the word shadowing task reveal a significant increase in the shadowers' use of rising intonation contours when repeating after the model talker in the Rise condition, relative to the shadowers' baseline. The shadowers' rising contours were also acoustically more similar to the model talker's rises in terms of f0 range and alignment in the shadowing block than in the baseline block. In the Fall condition, the female shadowers' falling contours were acoustically more similar to the model talker's falls in terms of f0 alignment in the shadowing block than in the baseline block. Together, these results suggest imitation of both phonological and phonetic features of intonation contours, as in previous work [9, 15, 16, 17, 21]. However, the results from the Fall condition, in which only phonetic imitation was observed, suggest that imitation of intonation contours need not be mediated by abstract phonological representations [cf. 22] and may reflect a phonetically-based perception-production link [1].

The evidence for imitation was more robust overall for the Rise condition than the Fall condition. Given that the primary difference between the two conditions was the shape of the intonation contour produced by the model talker, the salience of the contour may have affected the magnitude of imitation. Previous research [28, 29] has shown that more marked variants are imitated more than less marked variants and phonologically relevant variants or features may be imitated to a greater extent than variation that does not signal a change in meaning [10, 22]. Romera and Elordieta [13] found a similar asymmetry as in the current study in the imitation of L2 Majorcan Spanish intonation contours by L1 Peninsular Spanish speakers. Although both contours that they considered are falls, the intonation contour of interrogative sentences was imitated while that of declarative sentences was not. Romera and Elordieta [13] attributed this asymmetry to the relative perceptual salience of the interrogative contour. The interrogative fall has a relatively greater f0 range and steepness, making the tune more perceptually salient than that of the declarative tune.

Likewise, the difference in imitation of the rising and falling contours in the current study may reflect their relative salience. In American English, rises are associated with various functional meanings, such as listing, uncertainty, or "uptalk," whereas falls are associated with simple declaratives [30, 31]. The various meanings associated with rises in American English may result in the assignment of differential status to the contours, such that rises are perceived as marked relative to falls when they are produced on isolated words as in the current word shadowing task.

In conclusion, the observed patterns of implicit imitation of intonation contours in the word shadowing task in the current study demonstrate that both the phonological and phonetic features of intonation contours can be imitated without explicit instructions to imitate. However, the extent of imitation may be affected by the salience of the contour. In particular, more marked contours are imitated more robustly than less marked contours.

5. ACKNOWLEDGMENTS

We would like to thank Laura Beebe, Emily Behm, Elizabeth Bohinski, Sarah Chilson, Hali Clark,

McKenna Reeher, and Erika Shane for assistance with data collection and analysis. These data were collected at the Language Sciences Research Lab at the Center of Science and Industry in Columbus, OH.

6. REFERENCES

[1] Goldinger, S. D. 1998. Echoes of echoes? An episodic theory of lexical access. *Psych. Rev.* 105, 251-279.

[2] Pardo, J. S., Jay, I. C., Krauss, R. M. 2010. Conversational role influences speech imitation. *Atten. Percept. Psychophys*. 72, 2254-2264.

[3] Babel, M. 2012. Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *J. Phon.* 40, 177-189.

[4] Walker, A., Campbell-Kibler, K. 2015. Repeat what after whom? Exploring variable selectivity in a cross-dialectal shadowing task. *Front. Psychol.* 6, 546.

[5] Nielsen, K. 2011. Specificity and abstractness of VOT imitation. *J. Phon.* 39, 132-142.

[6] Shockley, K., Sabadini, L., Fowler, C. A. 2004. Imitation in shadowing words. *Percept. Psychophys.* 66, 422-429.

[7] Bonin, F., De Looze, C., Ghosh, S., Gilmartin, E., Vogel, C., Polychroniou, A., Salamin, H., Vinciarelli, A., Campbell, N. 2013. Investigating fine temporal dynamics of prosodic and lexical accommodation. In: *INTERSPEECH 2013, Lyon, France, August 25-29, 2013, Proceedings*. 539-543.

[8] Babel, M., Bulatov, D. 2012. The role of fundamental frequency in phonetic accommodation. *Lang. Speech* 55, 231-248.

[9] Sato, M., Grabski, K., Garnier, M., Granjon, L., Schwartz, J.-L., Nguyen, N. 2013. Converging toward a common speech code: Imitative and perceptuo-motor recalibration processes in speech production. *Front. Psychol.* 4, 422.

[10] Savino, M. 2017. The dynamics of prosodic adaptation between Italian conversational partners. In Botinis, A. (ed), *ExLing 2017, Heraklion, Crete, Greece, June 19-22, 2017, Proceedings*. 93-96.

[11] Pardo, J. S. 2006. On phonetic convergence during conversational interaction. *J. Acoust. Soc. Am.* 119, 2382-2393.

[12] Garnier, M., Lamalle, L., Sato, M. 2013. Neural correlates of phonetic convergence and speech imitation. *Front. Psychol.* 4, 600.

[13] Romera, M. Elordieta, G. 2013. Prosodic accommodation in language contact: Spanish intonation in Majorca. *Int. J. Sociol. Lang.* 221, 127-151.

[14] Lee, Y., Gordon, D. S., Parrell, B., Lee, S., Goldstein, L., Byrd, D. 2018. Articulatory, acoustic, and prosodic accommodation in a cooperative maze navigation task. *PLoS ONE* 13(8), e0201444.

[15] D'Imperio, M., Cavone, R., Petrone, C. 2014. Phonetic and phonological imitation of intonation in two varieties of Italian. *Front. Psychol.* 5, 1-10.

[16] Cole, J., Shattuck-Hufnagel, S. 2011. The phonology and phonetics of perceived prosody: What do listeners imitate? In: *INTERSPEECH 2011, Florence, Italy, August 27-31, 2011, Proceedings*. 969-972.

[17] Petrone, C., D'Alessandro, D., Falk, S. 2021. Working memory differences in prosodic imitation. *J. Phon.* 89, 1-17.

[18] Michelas, A., Nguyen, N. 2011. Uncovering the effect of imitation on tonal patterns of French Accentual Phrases. In: *INTERSPEECH 2011, Florence, Italy, August 27-31, 2011, Proceedings*. 973-976.

[19] Pardo, J. S., Jordan, K., Mallari, R., Scanlon, C., Lewandowski, E. 2013. Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. *J. Mem. Lang.* 69, 183-195.

[20] Pardo, J. S., Urmanche, A., Wilman, S., Wiener, J. 2017. Phonetic convergence across multiple measures and model talkers. *Atten. Percept. Psychophys.* 79, 637–659.

[21] German, J. S. 2012. Dialect adaptation and two dimensions of tune. In: *Speech Prosody 2012, Shanghai, China, May 22-25, 2012, Proceedings*. 430-433.

[22] Mitterer, H., Ernestus, M. 2008. The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition* 109, 168-173.

[23] Clopper, C. G., Carter, A. K., Dillon, C. M., Hernandez, L. R., Pisoni, D. B., Clarke, C. M., Harnsberger, J. D., Herman, R. 2002. The Indiana Speech Project: An overview of the development of a multi-talker multi-dialect speech corpus. In: *Research on Spoken Language Processing Progress Report No. 25*. Speech Research Laboratory, Indiana University, 367-380.

[24] Shue, Y.-L., Keating, P., Vicenik, C., Hu, K. 2011. VoiceSauce: A program for voice analysis. In: Lee, W. S., Zee, E. (eds), *ICPhS 2011, Hong Kong, August 17-21, 2011, Proceedings*. 1846-1849.

[25] Kawahara, H., Cheveigne, A., Patterson, R. D. 1998. An instantaneous-frequency-based pitch extraction method for high-quality speech transformation: Revised TEMPO in the STRAIGHT-suite. In: *ICSLP 98, Sydney, Australia, November 30-December 4, 1998, Proceedings*. 659-662.

[26] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *J. Stat. Softw*. 67, 1-48.

[27] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2017. lmerTest package: Tests in linear mixed effects models. *J. Stat. Softw.* 82, 1-26.

[28] Honorof, D. N., Weihing, J., Fowler, C. A. 2011. Articulatory events are imitated under rapid shadowing. *J. Phon.* 39, 18-38.

[29] Mitterer, H., Müsseler, J. 2013. Regional accent variation in the shadowing task: Evidence for a loose perception–action coupling in speech. *Atten. Percept. Psychophys.* 75, 557-575.

[30] Ladd, R. D. 1980. *The Structure of Intonational Meaning.* Indiana University Press.

[31] Warren, P. 2016. *Uptalk: The Phenomenon of Rising Intonation.* Cambridge University Press.