The phonology and phonetics of perceived prosody: What do listeners imitate?

Jennifer Cole[1], *Stefanie Shattuck-Hufnagel*[2]

[1] Department of Linguistics, University of Illinois at Urbana-Champaign
[2] Speech Communication Group, Research Laboratory of Electronics, MIT
jscole@illinois.edu, sshuf@mit.edu

Abstract

An imitation experiment tests the hypothesis that when asked to reproduce a spontaneously-spoken utterance that they hear, speakers imitate the prosody of the stimulus in its phonological structure more accurately than the phonetic details. Results suggest that speakers rarely distort the presence of a pitch accent or an intonational phrase boundary, but more often change the nature of the phonetic cues, e.g. the duration of a pause or the occurrence of irregular pitch periods associated with boundaries and accents in American English. These findings argue for an encoding of phonological prosodic structure that is separate from the phonetic cues that signal that structure.

Index Terms: prosody, spontaneous speech, spoken imitation, phonetics and phonology

1. Introduction

Recent work on the acoustic cues that speakers provide to the grammatical prosody of an utterance, i.e. its phrasal word groupings and its phrasal prominences, suggests that different speakers may select or emphasize different cues [1,2,3,4]. Furthermore, evidence is accumulating that listeners differ in their sensitivity to the acoustic cues to prosody, resulting in inter-listener differences in perception as a function of speaker [5]. Inter-speaker differences in prosody are especially salient in spontaneous speech, and contribute to the distinctive speech styles of individual speakers. One problem in investigating inter-speaker variation in prosody, however, is determining the source of prosodic differences that are identified in the speech signal. Inter-speaker differences in pitch contours, pausing, and other acoustic correlates of prosody can result from differences in the phonetic spell-out of phonological prosodic features, differences in the inventory of prosodic features (pitch accents and boundary tones), or differences in the assignment of prosodic features based on the syntactic, semantic or pragmatic properties of the utterance. Moreover, for studying prosody in spontaneous speech there is the added problem that utterances will differ in their lexical content, so differences between speakers may arise in part from differences in the lexical content of their utterances.

A potential solution to this problem lies in the use of imitation techniques, in which a listener is invited to reproduce a heard utterance that was produced by another speaker during a spontaneous conversation. In this condition all speakers produce the same target utterance (to an important extent), with the same pragmatic and discourse conditions (such as they are, in the production of an isolated utterance.) As a result, we can be certain that inter-speaker prosody differences that emerge are not due to the lexical content of their utterances, or to pragmatic or discourse conditions. Rather, variability in the prosody of imitated utterances across imitators (who are simultaneously listeners and speakers) may arise due to differences in (i) the imitator's perception of prosody in the original stimulus utterance; (ii) the influence of the syntactic form of the utterance on the imitator's production of prosody; or (iii) the idiosyncratic phonetic proclivities of the imitator in the expression of prosodic form.

This paper presents the results of one such study, designed to investigate which elements of prosodic form in spontaneous speech are reliably imitated by listeners. Using a speech imitation paradigm, we compare the prosodic form of imitated utterances to that of the original stimulus utterance at three levels: the **structural** level that defines the location of pitch accents and prosodic boundaries; the **featural** level that specifies the tonal features encoding the contrastive phonological categories of pitch accents and boundaries, and the level of detailed **phonetic cues** to prosodic structure and categories. More precisely, we test the hypothesis that imitators will more accurately reproduce the phonological structure and tonal features of the stimulus utterance, and somewhat less accurately reproduce the details of the phonetic realization of that structure. This hypothesis follows from the observation that speakers vary in the phonetic implementation of prosody [4,6]; for example, some speakers cue prominence primarily through duration, while others use intensity. Thus a speaker who perceives and faithfully imitates the prosodic structure and associated prosodic features of the stimulus may yet implement that prosodic form with a different set of phonetic cues than are present in the stimulus utterance. On the other hand, findings from speech shadowing studies lead us to predict that in some instances speakers may indeed imitate the detailed phonetic cues that accompany the prosodic structure and features in the stimulus utterance. This expectation follows from research showing that when speakers imitate an incoming utterance by shadowing (i.e. by reproducing it as quickly as possible, in real time, while they are hearing it), they often reproduce the broad prosodic form of the utterance [7] and imitate to some degree the phonetic detail of the speech they hear [8,9,10,11]. The extent to which imitations faithfully reproduce the phonetic detail conditioned by prosodic context will shed light on the separability of the phonological and phonetic encoding of prosody.

To further investigate the imitation of phonetic cues to prosody, we elicit imitations in series of three, in paced succession. If the phonetic cues to prosody present in the stimulus are imitated even in the final imitation in the series, that would suggest a tight bond between the phonological representation (prosodic structure and associated tonal features) and its phonetic cues, and would support the view that phonetic detail is encoded in memory with duration that exceeds the temporal limits of the short-term auditory buffer. On the other hand, it is also possible that later imitations will stray further from the stimulus at any or all levels of prosodic form (including prosodic structure, tonal features, and detailed phonetic cues), reflecting a greater influence of the imitator's own grammatical system [12].

2. Method

The imitation task was quite straightforward: each speaker heard the target utterance once, and then imitated it three times in succession.

Target utterances were drawn from a corpus of task-directed spontaneous speech, the American English Maptask Corpus, which was collected at the Speech Communication Group at MIT from 8 pairs of already-acquainted young (21-22 years) female speakers of an eastern dialect of American English [13]. In this cooperative task, the direction-giver conveys the detailed location of a path on one map to a colleague who has a similar map that lacks the path and also contains (unbeknownst to the speakers) slightly different landmarks. This task quickly engaged the speakers, who produced highly natural sounding speech as they discovered and worked out the differences.

From the resulting 16 dialogues, 8 utterances were extracted from mid-dialogue locations for each of 4 direction givers, resulting in 32 target utterances of moderate length (7-15 words, average length 11.5 words), as illustrated by the example utterance (1).

(1) *so you're gonna go between the mill wheel and the mountain…*

Participating speakers were 10 females (18-30 years old) who were students at the University of Illinois from the Midlands dialect area. They had no self-reported history of speech or hearing problems, and were paid $10 for their participation.

Target utterances were presented in auditory form to the participant, once only, in a quiet room. Speakers were instructed to "repeat the words and the way the utterance was said." This instruction was intended to elicit imitation of the lexical and syntactic content of the utterance, its prosodic form, and possibly also the speech rate. Subjects were intentionally not asked to *imitate* the target utterance, so as not to encourage a general impersonation of the stimulus speaker's voice. The speaker reproduced the utterance by speaking it aloud three times, pausing slightly between repetitions; no textual version of the sentence was presented.

The target utterances and their imitations were subjected to different kinds of prosodic labeling. Target utterances were independently transcribed by two experienced ToBI labelers using the ToBI standard [14,15] and disagreements were resolved via discussion. The pitch accents used to label phrasal prominence and the boundary labels marking ends of intermediate phrases (ip) and intonational phrases (IP) are shown in (2).

(2) Pitch accents: H*, L*, !H*, L+H*, L+!H*, L*+H, H+!H*
Phrase accents (ip): H-, L-
Boundary tones (IP): H%, L%

The imitated utterances were transcribed using the ToBI labeling criteria, but with a much reduced inventory of prosodic labels: 'A' for any kind of pitch accent, "B" for any kind of intonational boundary (ip or IP). The additional label 'a' was used to mark a word that sounded prominent, but which lacked any notable pitch movement, and 'b', for a location where a possible boundary was perceived with little pitch evidence. Two independent transcriptions were summed to obtain 3 levels of Accent and Boundary (2, 1, 0). The reduced labeling was chosen partly in the interest of time, and because comparison between the target and imitated utterance for the pitch accent and boundary *type* is being conducted through objective acoustic measures in our ongoing work.

Several different types of measures were used to test the hypotheses about the accuracy with which participants would reproduce the phonological structure of prosodic phrase boundary and accent (prominence) *vs.* the specific acoustic cues for those prosodic structures. First, we measured agreement in the phonological prosodic structure of the target and imitation utterances, in terms of the location of pitch accents and boundaries, assessing disagreement by the number of accents and boundaries of the target utterance that were "deleted" in the imitated utterance, and the number of novel prosodic elements that were "inserted" in the imitation. Second, we assessed the imitation of detailed non-intonational phonetic cues to prosody, in particular, boundary-related silence duration and voice quality variation in the form of irregular pitch periods.

We report preliminary results here, for 6 of the 10 participants and 16 of the 32 target utterances (eight target utterances each from two of the two selected Map Task speakers), and for the third (i.e. the final) repetition produced by each speaker. The phonological structure and phonetic accuracy measures are reported separately.

3. Results

A comparison of the pitch accent labels for the original stimulus utterances vs. the 3rd imitations from 6 subjects are shown in Figure 1, where Accented means both labelers agreed on the presence of an accent (2), MaybeAcc means that one labeler heard an accent and one did not (1) and Unaccented means that neither labeler heard an accent (0). As can be seen, the imitations generally followed the pattern of accented and unaccented syllables of the original utterance (Stimulus), with some variation from speaker to speaker. Counting accents that were labeled 2 and 1 together, the general pattern shows slightly more accents in the imitations relative to the stimulus utterances.

A comparison of the phrase boundary labels for the original target utterances and the 6 repetitions of each are shown in Figure 2. The similarity in the rates of boundary production is even more striking than that of the pitch accents.

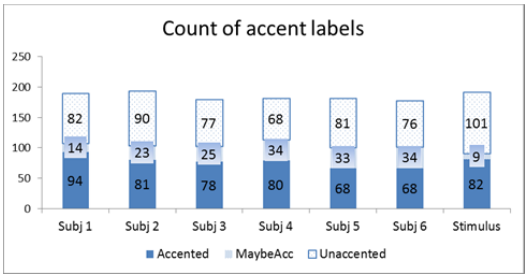

Figure 1: *Pitch accents in imitations and stimuli.*

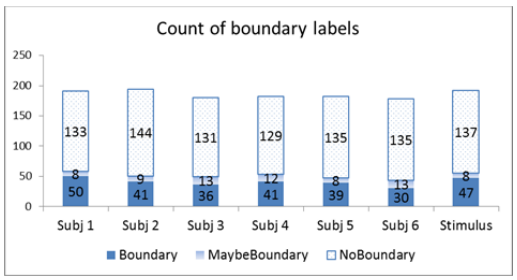

Figure 2: *Boundaries in imitations and stimuli.*

The agreement between stimulus and 3rd imitations for location of accents and boundaries is well above chance. Kappa statistics indicate agreement is "substantial": for Accent, the value was between 0.63 and 0.85, and for Boundary between 0.69 and 0.84. Another way of assessing the accuracy of speaker imitations is to evaluate the number of accents and boundaries that were inserted at locations where the original target lacked them, and omitted from locations where they were originally present. Figure 3 shows these data for accents and boundaries respectively. For both aspects of phonological structure, deletions from the original were rare, with between 1% and 3% of the total words in the stimulus utterances subject to accent deletion in the imitations produced by 6 speakers. In contrast, insertions were more common, affecting as many as 8% of the stimulus words. While both rates are low, the difference suggests that speakers occasionally supplement their imitation of the phonological structure of the target utterance prosody with additional phonological elements. In addition, there seems to be considerable variation across individual speakers.

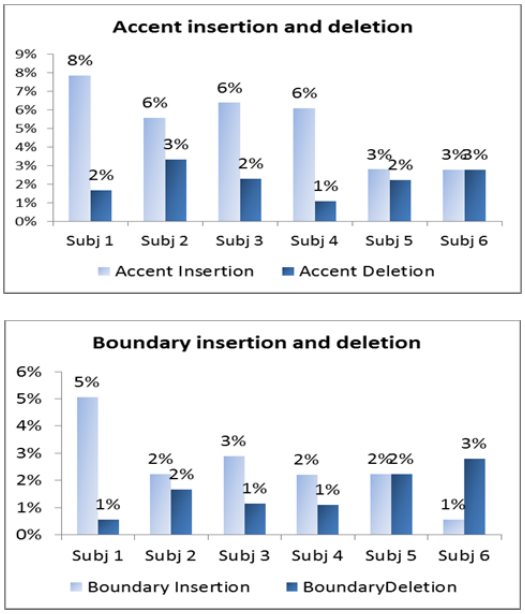

Figure 3: *Occurrence of accent insertion and deletion (top panel) and boundary insertion and deletion (bottom panel) calculated as the number of insertions/deletions over the total number of words in the set of target utterances.*

A further question about the accuracy with which speakers imitate the phonological structure of a target prosody concerns the distinction between nuclear and pre-nuclear accents. Nuclear pitch accents, defined as the final accent in an intonational phrase, appear to be more consistently associated with pragmatic focus [16]. If, as this claim suggests, nuclear accents are particularly significant for the listener in understanding an utterance, it is not implausible to predict that they will be more reliably imitated. Such a finding would also be consistent with earlier results showing greater inter-transcriber reliability for nuclear accents [5]. To test this possibility, we compared the rates of pitch accent discrepancy (both insertion and deletion) for nuclear vs. prenuclear accents. These results are shown in Figure 4. While all speakers inserted or deleted at least some accents, several failed to delete or insert any nuclear accents, and in general the insertion/deletion rates were lower for accents in nuclear position. This finding offers some support for the view that listeners (and speakers) pay particular attention to pitch accents in this semantically significant location.

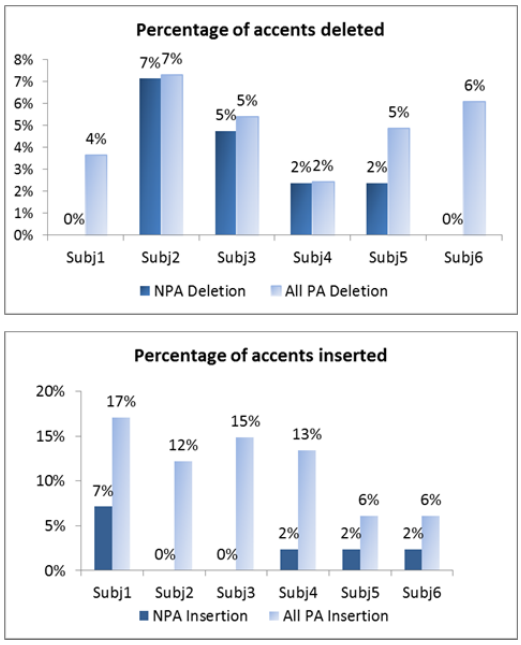

Figure 4: *Rate of Accent deletion (top panel) and insertion (bottom panel) calculated as the number of deletions/insertions over the total number of nuclear PAs (for NPA deletion/insertion), or over the number of all PAs in the target utterances (for All PA Deletion/Insertion).*

The results described so far indicate that speakers preserve the phonological encoding of the prosodic structure or form of an utterance that they imitate---but do they also preserve the details of how that phonological structure is implemented? The final hypothesis to be tested here is that speakers will be less accurate at reproducing the phonetic cues to prosodic structure than at reproducing the prosodic structure itself, in their imitations. Although we do not yet have summary quantitative data on this question, we have begun to examine two specific cues to prosodic boundaries; the duration of the pause (if any), and the presence of irregular pitch periods (IPP) [17]. The presence of IPP at phrase boundaries has been shown to vary significantly across speakers [1,18], and we reasoned that our 6 speakers might well prefer to implement this boundary cue in their own way, rather than following the lead of the target speaker, and perhaps similarly for pause duration.

Preliminary results of this analysis suggest that speakers do not perfectly reproduce the details of the phonetic implementation produced by the target speaker. For example, comparing pause duration after *peak* in the utterance "*Um, you're gonna be standing at the peak of the mountain, on the Canadian Paradise*" shows that the target utterance had no pause (although a prosodic boundary was marked by other cues, such as final lengthening and intonation), while the imitations of 4 different speakers had pauses of 42, 124, 0 and 88 ms (Table 1). For the presence of IPP in the same utterance, the original speaker produced irregular periods on *um* and *standing*, while only some of 4 speakers did so.

4. Discussion

The results of this analysis are consistent with the view that listeners perceive the basic prosodic structure of an utterance, i.e. the presence and location of pitch accents and intonational

Table 1. *Pause duration (ms) and irregular pitch periods (IPP) in the utterance* um] you're gonna be standing at the peak] of the mountain] on the Canadian *Paradise (prosodic boundaries at right brackets) for stimulus utterance and imitations from four subjects (S1-S4). Words with no following pause or no IPP are omitted from table.*

	um	*you're*	*standing*	*peak*	*mount.*	*Par.*
Stim.	114; ipp	--	ipp	0	0	--
S1	0	--	--	0	42; ipp	--
S2	0	ipp	--	68	124	--
S3	0	ipp	ipp	45	0; ipp	ipp
S4	95	--	--	68	88	--

phrase boundaries, and reproduce this structure even in their third imitations of the target. These aspects of the prosody are rarely deleted, although some insertions occur at this level. One might ask whether this consistency arises because speakers generate the same prosodic structure as the original speaker did, based on the same syntactic, semantic and pragmatic factors. However, this seems unlikely, since it is well known that syntactic structure severely underdetermines intonational phrase boundary locations [19], and other relevant factors that were known to the original speaker, such as pragmatic and discourse context, were not available to the participants in the experiment. On the other hand, we have evidence that speakers producing imitations are not simply 'parroting' the prosody of the original speaker, since insertions and deletions of basic elements of the prosodic phonology do occur. Thus we infer that imitation involves the generation of prosodic structure to some extent.

In addition, we have preliminary evidence that the details of the non-intonational phonetic cues to prosodic structure may not be as reliably reproduced in the imitations as the basic prosodic structure is. If supported by further analysis, this finding would be consistent with the claim that speakers in a shadowing task reproduce the phonological form of the target utterance, but not its phonetic detail, in the domain of segmental cues [20]. However, such a finding would appear to be at odds with reports of accurate shadowing of phonetic detail [21, 22]. Of course, such discrepancies might arise from the fact that different cues were analysed here, or from the use of different tasks (online shadowing vs. delayed imitation).

It is clear from these results that ordinary listeners can in general reliably imitate the basic prosodic structure of an utterance, with certain aspects of the structure imitated somewhat more reliably than others; for example, boundaries more reliably than accents, and nuclear accents more reliably than pre-nuclear accents.

The imitation method also provides a tool for the study of individual speaker signatures in the phonetic implementation of prosodic phonology. Although listeners can apparently accommodate to such variation, they do not necessarily align their production habits with those of the speaker they listen to, at least as long as that speaker is not present in an interaction.

5. Conclusions

Results of this utterance imitation study support the view that listeners generally perceive the basic prosodic structure of an utterance they hear, and reproduce it with a high degree of accuracy (although not perfectly) in their imitations. Intonational phrase boundaries are particularly reliably reproduced, and nuclear accents are distorted less than prenuclear accents. In addition, the imitation method provides an index of how accurately the details of the phonetic cues to prosodic structure are perceived, remembered and reproduced. Preliminary results support the hypothesis that phonological structure is encoded separately from the phonetic cues that signal it, and that these cues vary from one speaker to another.

6. Acknowledgements

We thank Dayna Cueva-Alegria, JC's research assistant.

7. References

[1] Dilley, L., Shattuck-Hufnagel, S., and Ostendorf, M., Glottalization of vowel-initial syllables as a function of prosodic structure. *Journal of Phonetics* 24, 423-444, 1996.

[2] Peppé, S., Maxim, J., and Wells, B., Prosodic variation in British English. *Language and Speech* 43: 309-334, 2000.

[3] Grabe, E., Variation adds to prosodic typology. *Proc. Speech Prosody*, Aix-en-Provence, France, 2002.

[4] Mo, Y., *Prosody Production and Perception with Conversational Speech,* Ph.D. thesis, Univ. of Illinois, 2010.

[5] Mo, Y., Cole, J. and Lee, E-K., Naive listeners' prominence and boundary perception. *Proc. Speech Prosody,* Campinas, Brazil, 2008.

[6] Yoon, T., Speaker consistency in the realization of prosodic prominence in the Boston University Radio Speech Corpus, *Proc. Speech Prosody,* Chicago, IL, 2010.

[7] Marslen-Wilson, W.D., Speech shadowing and speech comprehension. *Speech Communication* 4, 55-73, 1985

[8] Goldinger, S., Echoes of Echoes? An Episodic Theory of Lexical Access. *Psych. Rev.* 105: 251-279, 1998.

[9] Shockley, K., Sabadini, L., and Fowler, C., Imitation in shadowing words. *Attention, Perception & Psychophysics* 66: 422-429, 2004.

[10] Pardo, J., On phonetic convergence during conversational interaction. *JASA* 119: 2382-2393, 2006.

[11] Nye, P. & Fowler, C., Shadowing latency and imitation: The effect of familiarity with the phonetic patterning of English. *Jn Phonetics,* 31: 63-79, 2003.

[12] Braun, B., Kochanski, G., Grabe, E., and Rosner, B., Evidence of attractors in English intonation. *JASA* 119: 4006-4015, 2006.

[13] Shattuck-Hufnagel, S. and Veilleux, N.M., The robustness of acoustic landmarks in spontaneous speech. *ICPhS,* Saarbrueken, 2007.

[14] Silverman, K., Beckman, M., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J., ToBI: a standard for labeling English Prosody. *ICSLP,* 1992.

[15] ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-911-transcribing-prosodic-structure-of-spoken-utterances-with-tobi-january-iap-2006/

[16] Calhoun, S., *Information Structure and the Prosodic Structure of English: A Probabilistic Relationship,* PhD thesis, Univ. Edinburgh, 2006.

[17] Slifka, J., Some physiological correlates to regular and irregular phonation at the end of an utterance, *Journal of Voice* 20: 171-186, 2006.

[18] Bohm, T. and Shattuck-Hufnagel, S., Do listeners store in memory a speaker's habitual utterance-final phonation type? *Phonetica* 66: 150-168, 2009.

[19] Cole, J., Mo, Y., and Baek, S. The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech. *Language and Cognitive Processes,* 25: 1141-1177, 2010.

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

[21] Fowler, C. A., Brown, J. M., Sabadini, L., & Weihing, J., Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Jn Memory and Language,* 49: 296–314, 2003.

[22] Brouwer, S., Mitterer, H., and Heuttig, F. Shadowing reduced speech and alignment. JASA 128: EL32-EL37, 2010.