Are listeners sensitive to the phonological form of prosody or its phonetic encoding?

Jennifer Cole
University of Illinois at Urbana-Champaign, USA

Stefanie Shattuck-Hufnagel
Massachusetts Institute of Technology , USA

Hansjörg Mixdorff
BHT University of Applied Sciences, Berlin, Germany

Memory for spoken words includes memory of the speaker's voice and other contextually specified phonetic properties, but there is also evidence that memory encoding is in terms of abstract phonological segments or features.

(Goldinger 1988; Church & Schacter 1994)
(Vitevitch & Luce 1998, 2005; Cutler et al. 2010)

Broad Research Question: What are the roles of phonological and phonetic form in the perception, memory encoding, and production of spoken utterances?

Our focus: the phonology and phonetics of prosody (prominence and phrasing) in American English

Hypothesis 1: listeners are sensitive to the phonetic detail of the prosody of spoken utterances, but what they store to guide future production is the phonological form.

Hypothesis 2: when asked to reproduce an utterance they have heard, speakers will replicate the broad phonological form of prosody, but using their own phonetic exponents.

Background: *imitation of segmental cues*

Speaker variability. Speakers vary in the acoustic cues they provide to the grammatical prosody of an utterance—its phrasal word groupings and its phrasal prominences.
(Dilley et al. 1996; Grabe 2002; Mo 2010; Yoon 2010)

Listener sensitivity. Listeners are sensitive to the acoustic cues to prosody, and adapt their perceptual response to the cue pattern of an individual speaker. (Mo et al. 2008)

Perception – production link. The phonetic detail perceived by the listener can influence their subsequent production. Evidence from speech shadowing studies shows that speakers imitate sub-phonemic phonetic detail, at least when it is linguistically relevant.
(Fowler et al. 2003; Mitterer & Ernestus 2008; Shockley et al. 2004)

Primacy of the canonical phonological form. Though listeners may perceive phonetic detail of a specific stimulus signal, in shadowing tasks speakers show a bias towards canonical forms. Canonical forms are shadowed at faster response times than reduced forms, and reduced forms are often restored to a canonical, full form in the imitated production.
(Brouwer et al. 2010)

Global convergence in prosodic detail. The influence of perceived phonetic detail on subsequent production is also observed in natural discourse settings, as speakers converge on the phonetic patterns of their interlocutor. Convergence is judged perceptually based on global similarity. (Pardo 2006; Kim 2011)

What about prosody?
- Are speaker-specific phonetic cues to prosodic features imitated?
- Is there prosodic convergence with spontaneous speech?

Methods

We elicited imitations of pre-recorded, spontaneously produced utterances. Subjects heard each target utterance once and imitated it three times in succession.

Stimuli

- 32 utterances of spontaneous speech (7-15 words each) were extracted from 4 young adult, female speakers (Eastern dialect) from the American English Maptask corpus.
- Pairs of Maptask speakers engaged in a cooperative task, with one speaker acting as a direction-giver conveying the detailed location of a path on one map to their partner who has a similar map that differs in some landmarks.

Sample utterance:
"so you're gonna go between the mill wheel and the mountain"

Imitators

- Imitators were 10 young adult females (18-30 years old), students at the University of Illinois and native speakers of American English (Midland dialect), who were paid $10. *Data from four (up to six) subjects are presented here.*

Task

- Target utterances were presented through headphones, with no text presentation.
- Imitators were instructed to listen to each utterance and then to *"repeat the words and the way the utterance was said."*
- Goal: to elicit imitation of the lexical and syntactic content, and the prosodic form of spontaneous speech utterances
- Imitators repeated each utterance three times in succession, pausing briefly between repetitions. *Data from the third repetition are presented here.*

Phonological prosodic labeling

- Target utterances were independently transcribed by two of the authors (JC, SSH) using ToBI, with differences resolved through discussion to obtain a consensus ToBI labelling.

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

- Imitated utterances were independently transcribed by the same transcribers using ToBI criteria, but with a reduced label inventory:
  - **"A"** for prominence marked by salient pitch movement
  - **"a"** for perceived prominence with no pitch movement
  - **"B"** for prosodic boundary (ip or IP) marked with pitch
  - **"b"** for perceived boundary not marked by pitch
- Independent transcriptions were summed to obtain three levels of Accent and Boundary: 2, 1, 0.

Acoustic correlates of prosody

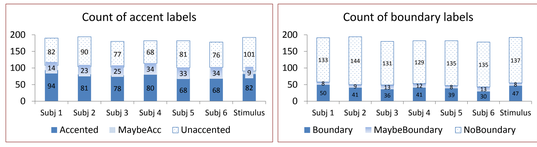
Measured for both target and imitated utterances:

- **Glottalization**, as a cue to prosodic boundaries was manually labeled by one author (SSH) based on irregular pitch periods visually observed in the waveform.
- **Syllable durations**, expected to be longer in phrase-final and phrase-initial positions, and in accented syllables
- **F0 cues** to pitch accent were extracted from the parameters of the Fukisaki model (Mixdorff 2000): accent command amplitude (Aa), accent type (rise *vs.* fall), accent alignment

Results

Descriptive statistics

Total number of Accent and Boundary labels over all utterances (approx. 180 words), for each speaker and for stimulus (target) utterances..



Statistical analysis of target-imitation similarity

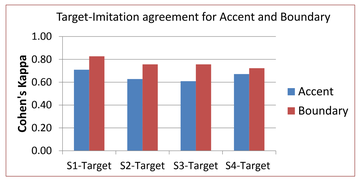
Target-Imitation. Cohen's Kappa statistic measures the agreement between the target utterance and each imitation. *The analysis here is limited to the third of the three successive imitations subjects produced for each target utterance.*

$$\text{Cohen's Kappa: } \frac{Agree_O - Agree_E}{1 - Agree_E}$$

$Agree_O$ is observed agreement over all labels; $Agree_E$ expected agreement (chance). Kappa scores of .41–.60 can be taken as moderate agreement, .61–.80 as substantial, and .81–1 as almost perfect agreement.

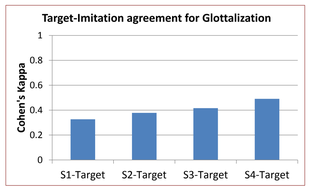
Agreement in phonological prosodic structure

Imitations show substantial agreement with target utterances for the location of Accents and Boundaries

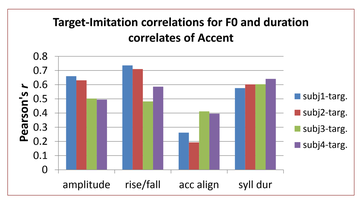


Agreement in acoustic-phonetic correlates of prosody

Imitations show fair to moderate agreement with target utterances for the occurrence of glottalization (per syllable)

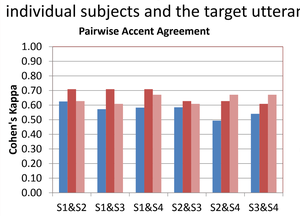


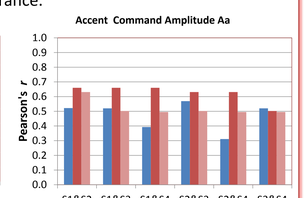
There are significant though moderate correlations between target and imitated utterances for F0 parameters and duration

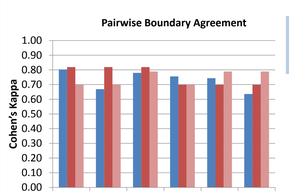


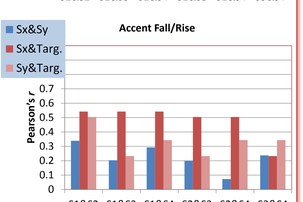
Pairwise agreement between subjects

Could agreement between target and imitation be driven by a "default" prosody, determined by the lexical and syntactic context? If so, agreement between pairs of subjects should be comparable to agreement between individual subjects and the target utterance.









Discussion

Results are consistent with Hypothesis 1:

Although listeners are sensitive to sub-phonemic detail in an utterance, they store the phonological structure

Results are also consistent with Hypothesis 2:

Speakers use that phonological structure to guide imitations produced with their own phonetic exponents.

What is the role of subphonemic detail in this task?

- provides cues to prosodic structure, speaker specificity, etc.
- unlikely to lead to 're-tuning' of phonemic category boundaries (Cutler 2010) because stimuli include only a few tokens from each speaker.

Significance of these results

- Although listeners are sensitive to and influenced by sub-phonemic detail, and shadowers can reproduce it, our imitators more reliably reproduce the linguistically-significant contrastive phonological structures of prosodic form.

Do shadowing and imitation tasks tap into different aspects of the perception-storage-production process?

Questions for Future Work

- Do these findings hold for other phonetic exponents? -e.g., silence duration, F0 alignment
- How consistent and how detailed are individual speakers in their use of phonetic exponents? -e.g. placement of irregular pitch periods within a syllable
- Are individual speakers more consistent for some exponents than for others?
- Are there significant differences from Imitation 1 to Imitation 3, as the memory of the original stimulus fades?

References

Brouwer, S. Mitterer, H., Huettig, F. 2010. Shadowing reduced speech and alignment. *JASA*, 128(1), EL32-EL37.
Church, B. A. and Schacter, D. L. (1994). Perceptual specificity of auditory priming: implicit memory for voice, intonation, and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory,* and *Cognition* 20: 521–33.
Cutler, A., Eisner, F., McQueen, J., and Norris, D. 2010. Coping with speaker-related variation via abstract phonemic categories. In Fougeron, C., D'Imperio, M., Kühnert, B., and Vallée, N. (eds.), *Papers in Laboratory Phonology X*, pp. 91-111. Mouton de Gruyter, Berlin, Germany.
Dilley, L., Shattuck-Hufnagel, S., and Ostendorf, M. 1996. Glottalization of vowel-initial syllables as a function of prosodic structure. *J Phonetics*, Vol. 24, 423-444.
Fowler, C. A., Brown, J., Sabadini, L., and Weihing, J. 2003. Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks, *J. Memory and Language.* 49, 396–413.
Goldinger, S. D. 1998. Echoes of echoes? An episodic theory of lexical access, *Psychol. Rev.* 105, 251–279.
Grabe, E. 2002. Variation adds to prosodic typology. In B.Bel and I. Marlin (eds), *Proc. Speech Prosody*, Aix-en-Provence, 127-132.
Fowler, C. A., Brown, J. M., Sabadini, L., & Weihing, J. 2003. Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *J Memory and Language*, 49, 296–314.
Kim, M., Horton, W., and Bradlow, A. 2011. Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. *Laboratory Phonology 2*, 125-156.
Mitterer, H., and Ernestus, M. 2008. The link between speech perception and production is phonological and abstract: Evidence from the shadowing task, *Cognition* 109, 168–173.
Mixdorff, H. 2000. A novel approach to the fully automatic extraction of Fujisaki model parameters. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 2000, vol. 3. Istanbul, 1281-1284.
Mo, Y. 2011. *Prosody production and perception with conversational speech.* Doctoral dissertation, Linguistics, Univ. of Illinois.
Mo, Y., Cole, J., Lee, E-K. 2008. Naïve listeners' prominence and boundary perception. *Proc. Speech Prosody 2008*, pp. 735-738. Campinas, Brazil.
Pardo, J. S. 2006. On phonetic convergence during conversational interaction. *JASA* 119, 2382–2393.
Shockley, K., Sabadini, L., & Fowler, C. A. 2004. Imitation in shadowing words. *Perception & Psychophysics*, 66, 422–429.
Vitevitch, M. S., & Luce, P. A. 1998. When words compete: Levels of processing in spoken word perception. *Psychological Science*, 9, 325–329.
Vitevitch, M. S., & Luce, P. A. 2005. Increases in phonotactic probability facilitate spoken nonword repetition. *J Memory and Language*, 52 (2), 193-204.
Yoon, Tae-Jin (2010): "Capturing inter-speaker invariance using statistical measures of rhythm", *Proc. Speech Prosody 2010.*