Specificity and Generalizability of Spontaneous Phonetic Imitation

Kuniko Y. Nielsen

Department of Linguistics
University of California, Los Angeles, USA
kuniko@humnet.ucla.edu

ABSTRACT

The imitation paradigm [1, 2], in which subjects' speech is compared before and after they are exposed to target speech, has shown that subjects shift their production in the direction of the target, indicating the use of episodic traces in speech perception as well as the close tie between speech perception and production. By using this paradigm, the current study aims to investigate the psychological reality of three levels of linguistic unit (i.e., word, phoneme, and feature). An experiment was designed to test whether spontaneous phonetic imitation can be *generalized* from words across (a) new words which share the same initial phoneme, and (b) new words with a new phoneme falling in the same natural class (sharing a feature); and also whether word-level specificity can be obtained through physical measurements instead of perceptual assessments. The feature manipulated in the experiment was aspiration, or [+spread glottis], on the phonemes /p/ and /k/.

The results showed that subjects produced significantly longer VOTs after they were exposed to target speech with longer VOTs, replicating [2] in a *non-shadowing* paradigm. Furthermore, the modeled feature (increased aspiration) was generalized to new instances of /p/ (i.e., in new words) as well as to the new segment /k/. At the same time, subjects' post-exposure VOT was significantly longer for those items that were in the target speech than items which they had not previously heard. These results, taken together, indicate that speakers possess both sub-phonemic and word-level representations.

Index Terms: speech perception, speech production

1. INTRODUCTION

Recent studies have shown that traces of episodic memory are retained and used in speech perception [3], and that both speech perception [4, 5] and production [6, 7] are more plastic than previously considered. The imitation paradigm, in which subjects' speech is compared before and after they are exposed to target speech, has shown that speakers shift their productions in the direction of what they just heard. For example, Goldinger [1] showed that subjects shifted their own F0 after listening to speech with artificially manipulated F0. His result also revealed a word-specific advantage of imitation: larger imitation effects were observed among low-frequency words than high-frequency words. This is as predicted by exemplar-based theories, because the smaller the number of exemplars associated with a given word, the larger the weight of each new exemplar. Shockley et al. [2] extended Goldinger's work by showing a significant Voice-Onset-Time (VOT) imitation effect for voiceless stops with artificially extended VOTs obtaining physical measurements of VOT.

These results demonstrate not only listeners' sensitivity to variations in global phonetic dimensions such as overall pitch range, but also sensitivity to the fine phonetic detail of a single segment such as degree of aspiration. Although Goldinger's results show evidence for word-size representations, they do not reveal whether sub-lexical units were also influenced by the imitation effect. That is because the post-exposure productions were elicited in the form of shadowing (= immediate repetition), and thus the listening and production lists had to be identical. The present study extended the earlier studies by using a *non-shadowing* task, which lets the listening (=target) and production lists differ; thus unheard words can be introduced into the production list. This allows us to test the generalizability of the imitation effect to sub-lexical units. Many linguistic theories (e.g., [8]) assume three levels of representations: lexical (=word), phonemic, and sub-phonemic (= feature or gesture). In order to test these assumptions through spontaneous phonetic imitation, the following questions were asked:

1) Generalizability to new stimuli
   a) Will there be generalization to the same phoneme in new (unheard) words?
   b) Will there be generalization to the same feature in new phonemes?

If we observe sub-lexical generalization at the phoneme level but not the sub-phonemic (feature) level, it will provide support for phoneme-size representations. If we observe generalization at both phoneme and sub-phonemic levels independently, it will provide support for phoneme and feature-size representations.

2) Word-specific advantage
   a) Will there be a larger imitation effect for words in the listening list?
   b) Will there be a larger imitation effect for words with lower lexical frequency?

The exemplar view predicts a stronger specificity for more recently experienced words. So we would expect a larger

imitation effect for words which subjects heard in the experiment. Also, as already shown in Goldinger, the exemplar view predicts larger specificity for low-frequency words than for high-frequency words.

2. METHOD

Participants. Seventeen native speakers of American English with normal hearing served as subjects for this experiment. They were recruited from the UCLA undergraduate population, and included 9 females and 8 males. They received course credit for participating.

Stimuli. The production list consisted of 150 English words. Among them, 100 were words beginning with /p/ (80 target words: 40 high-frequency words and 40 low-frequency words which were played in the study phase, and an additional 20 low-frequency words which were not played during the listening phase), and 20 were low-frequency words beginning with /k/. The remaining 30 words began with sonorants and served as fillers. The listening list consisted of 120 English words, including 80 target words from the production list (40 high-frequency words and 40 low-frequency words beginning with /p/), and 40 filler words beginning with sonorants. The lexical frequency was determined from both Kŭcera & Francis [9] and CELEX2 [10]: the threshold for low-frequency words was 5 (per million) and 300, and that for high-frequency words was 50 and 1000, respectively. The phonological neighborhood density and syllable length were controlled between the two frequency groups. All the words had equally high familiarity (> 6.0 on the 7-point Hoosier Mental Lexicon scale) [11]. All the target words had initial stress, and there were no onset clusters.

A phonetically trained male American English speaker recorded the 120 words in the listening list. The speaker first produced the words in the list normally, and then he produced the target words (words beginning with /p/) with extra aspiration. The VOT for the normally produced initial /p/ was measured, and was spliced with the initial part of hyper-aspirated tokens using PCquirer (Scicon R&D, CA) so that the resulting tokens have VOT extended by 40ms. The extended tokens had VOT of 113.26 ms on average (SD=10.82). This splicing method was chosen, as opposed to extending the middle part of VOT, in order to maximally preserve natural formant transitions.

Procedure. The experiment used a slightly modified version of the imitation paradigm [1, 2], in that a warm-up reading phase was added at the beginning to avoid possible hyper-articulation in the test reading due to first exposure. The stimuli were presented using Psyscope 1.2.5 [12]. Each subject was seated in front of a computer in a sound booth. Each session was divided into 4 blocks: 1) warm-up, 2) baseline, 3) listening, and 4) test. In the warm-up block, the words were presented, one at a time, on a computer screen every 2 seconds. The subjects were instructed to read the words silently without pronouncing them. In the baseline block, the subjects were instructed to "identify the word you see by speaking it into the microphone." In the listening block, using headphones, the subjects were exposed to two repetitions of the 120 spoken word tokens (80 target words and 40 filler words). There was no additional task during this block. The test block was exactly the same as the baseline block. Across the four blocks, the words were presented in random order for each subject. The subjects' tokens were digitally recorded into a computer and VOTs were measured using both waveforms and spectrograms. Unlike in previous studies, there was no perceptual assessment (i.e., AXB testing) of the baseline versus test productions.

3. RESULTS

Within-subject factors in this study were:

Type of Production: (Baseline vs. Test)
Lexical Frequency: (High vs. Low)
Word Specificity: (Target vs. Novel Items)
Segment: (p/ vs. k/)

Repeated-measures ANOVA analysis with two within-subjects factors (Type of Production and Lexical Frequency) revealed a significant difference between baseline vs. test productions ($F(1,16)= 9.167$, $p<.01$*), while the difference between high and low frequency groups did not reach significance ($F(1,16)= 4.238$, $p<.056$). The interaction between the two factors was not significant ($F(1,16)= .356$, $p>.1$).

Another repeated-measures ANOVA analysis with two within-subjects factors (Type of Production and Word Specificity [= target vs. novel stimuli]) showed a significant difference for both Type of Production ($F(1,16)=8.492$, $p<.01$*), and Word Specificity ($F(1,16)=10.083$, $p<.01$*). However, the interaction between the two factors was not significant ($F(1,16)=2.261$, $p>.1$).

Lastly, in order to see how the imitation effect is generalized to new stimuli, a repeated measures ANOVA with two within-subjects factors (Type of Production and Segment) were performed. Note that neither group of words was played in the listening block. Similar to the earlier tests including items that were played in the listening block, there was a significant difference between pre- and post-exposure productions ($F(1,16)=11.089$, $p<.01$*). As expected, there was a significant difference between /p/ and /k/ ($F(1,16)=234.09$, $p<.001$*), while there was no interaction between the two factors ($F(1,16)=0.275$, $p>.1$.)

Table 1 shows the medians, means, standard deviations and standard errors in VOT (ms) by stimulus types. As you can see, the standard deviations are very large in general, due to the individual variability of VOT. On the other hand, the means of standard errors are quite small, which shows that the subjects' shifts in their production (= imitation) were rather consistent.

4. DISCUSSION

Our results revealed a statistically significant difference between the baseline vs. test productions. As you can see in Figure 1, test-productions (lighter bars) show consistently longer VOTs than baseline productions (darker bars),

revealing that the VOT imitation effect is present even when the task involves elicitation-style production. This result is consistent with previous studies [1, 2] as well as the episodic view of speech perception. About eight minutes after they heard the target speech, subjects sustained its detailed surface (phonetic) information (i.e., extended aspiration).

Table 1: Summary of Results

Stimuli Type	Production	VOT Median (ms)	VOT Mean (ms)	Std. Deviation	Std. Error. Mean
Target P Low	B	60.130	67.810	19.2306	4.6641
	T	79.900	76.650	19.1954	4.6556
Target P High	B	59.910	66.856	18.6443	4.5219
	T	78.880	74.867	19.0002	4.6082
Novel P Low	B	61.314	66.292	18.5977	4.5106
	T	71.890	72.644	17.1517	4.1599
Novel K Low	B	72.000	77.748	17.9456	4.3525
	T	83.300	83.390	15.3913	3.7329

Our data also showed that the imitation effect was generalized to novel stimuli that subjects did not hear during the listening block (see *Novel* types in Figure 1). Compared with their baseline, subjects produced significantly longer VOT in the test block even for the novel words with initial /p/. This result indicates that the locus of spontaneous phonetic "imitation" is not word-specific, and that subjects imitated something smaller than a word.

Probably the most important finding of the current study is that the imitation effect was also generalized to a new phoneme /k/, which shares the manipulated feature [+spread glottis]. This result indicates that subjects imitated a unit that is smaller than the phoneme, suggesting subjects' knowledge of sub-phonemic representation. Many linguistic theories assumes that words are made up of discrete speech sounds (=segments) that are themselves complexes of features [8]. Support for this notion has traditionally been provided by phonological alternations and phonotactic constraints. Recent experimental work [13] suggests that speakers possess knowledge of phonological structure: in this study, subjects were able to learn phonotactic constraints at two levels of representations (segment and feature) through exposure to a set of nonwords. The current study provides additional support for the sub-phonemic assumption through spontaneous phonetic imitation. On the other hand, there was no interaction ($F<1$, $p>.1$) between the tested segment (i.e., /p/ and /k/) and Type of Production (baseline vs. target), showing that the amount of imitation for the two groups was the same. Thus although the imitation was generalized into novel items with initial /p/, our result for phoneme-level representation remains inconclusive. Note also that these results only confirm sub-phonemic representations, but not necessarily distinctive features as such. For example, it is entirely possible that the imitated unit is a gesture instead of a feature. These two theoretically contrastive views are in fact indistinguishable in the current study.

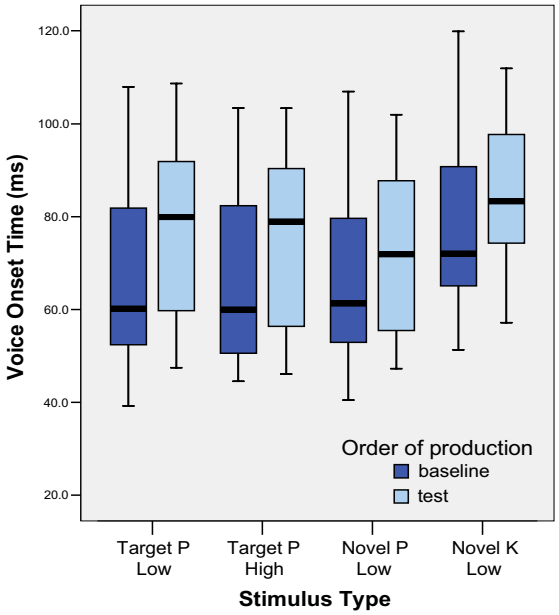

Figure 1: Imitation effect (in VOT) plotted across four types of stimuli.

In order to replicate the effect of lexical frequency [1] in a non-shadowing VOT paradigm, the current study carefully controlled frequency as an independent variable. Contrary to our expectation, the difference between the two frequency groups did not reach significance, and there was no interaction between the imitation effect and lexical frequency. This unexpected result may be attributed to the "warm-up" block, which did not exist in the original imitation paradigm [1, 2]. As shown in [14], a lexical frequency effect could easily disappear in such a repeated-sampling procedure. Although this modification was done in an attempt to eliminate/minimize some hyper-articulation (for low-frequency items) observed in our pilot study, it did increase the number of exposures to the target words. Given that our data appear to show the expected trend, a study with more statistical power, or, one without a warm-up block may detect the effect.

Although the effect of lexical frequency was not found, the result revealed a word-specific advantage in the target vs. novel words comparison, which was also predicted by the exemplar view. The target stimuli that subjects were exposed to in the listening block showed a significantly stronger imitation effect than the novel stimuli. This result provides support for the concept of word-level phonetic representation.

This word-specific effect argues against interpreting the imitation as due to global changes in speech style: if the change is due to global aspects of speech, such as changes in register, we would not expect to see the significant

difference between the target vs. novel words comparison. A further argument against such an interpretation comes from our post-hoc analysis of whole-word duration. If the effect is due to episodic memory or rule-learning, only the manipulated variable (in this case, VOT) should be affected. On the other hand, if the change is due to more global aspects of speech, we would expect to see changes in other variables. For this reason, the whole-word duration of the low-frequency target words was measured from randomly chosen 8 subjects' data. Unlike VOT, there was no significant difference between baseline and test ($F<1$, $p > .1$) productions. Given these results, it is unlikely that global aspects of speech are solely responsible for the spontaneous phonetic imitation observed in this study.

5. CONCLUSION

In order to see if there is experimental support for the structures assumed by many linguistic theories, a non-shadowing spontaneous phonetic imitation experiment was conducted which tests 1) generalizability of phonetic imitation to new instances which share (a) the same initial phoneme, or (b) the same feature, and 2) the word-specific advantage predicted by exemplar view. As expected, the results revealed a significant effect of phonetic imitation in a non-shadowing paradigm: subjects produced significantly longer VOTs after they were exposed to the target speech than their baseline productions recorded prior to the exposure. Furthermore, the results showed that the modeled feature [+ spread glottis] was generalized to new words (with initial /p/) as well as to a new segment (/k/). This result indicates that the subjects possess knowledge on sub-phonemic structure, supporting the traditional assumption in linguistics. Although the expected effect of lexical frequency was inconclusive in our data, we found a word-specific advantage in the comparison of target/novel items.

This study showed that speakers are sensitive to, and remember, sub-segmental detail, which therefore must be represented in some way. At the same time, it confirmed the word-specific advantage of phonetic imitation predicted by the exemplar view. These findings are compatible with models of spoken word recognition with a sub-phonemic level of representation, and are also parallel to proposed exemplar models of speech production [15]. The results of the current study thus call for a linguistically informed model of speech perception, which incorporates both sub-segmental and word-level representations.

6. ACKNOWLEDGEMENTS

The research was supported by NSF Grant BCS-0547578 (PI: Patricia Keating) and a UCLA Dissertation Year Fellowship. The author would like to thank the audience at RIKEN BSI Forum as well as LSA Albuquerque for their comments.
Correspondence concerning this article should be addressed to Kuniko Nielsen, Department of Linguistics, UCLA (e-mail: kuniko@humnet.ucla.edu).

7. REFERENCES

[1] Goldinger, S. D. (1998). Echoes of Echoes? An Episodic Theory of Lexical Access. *Psychological Review,* 105 (2), 251-279.

[2] Shockley, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception & Psychophysics*, 66 (3), 422-429.

[3] Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989) Some effect of talker variability on spoken word recognition. *Journal of the Acoustic Society of America*, 85, 365-378.

[4] Norris, D., McQueen, J. M., and Cutler, A. (2003). Perceptual learning in speech. *Cognit. Psychol.*, vol. 47, no. 2, pp. 204-238.

[5] Clark, C. M., and Luce, P.A. (2005) Perceptual adaptation to speaker characteristics: VOT boundaries in stop voicing categorization. *Proceedings of ISCA Workshop on Plasticity in Speech Perception (PSP2005)*; London, UK.

[6] Wright, R. (2004). Factors of lexical competition in vowel articulation. In J. Local, J., R.Ogden, and R. Temple (eds.). *Papers in Laboratory Phonology VI.* Cambridge: Cambridge University Press.

[7] Hay, J. B. (2000) *Causes and Consequences of Word Structure*. Ph.D dissertation, Northwestern University.

[8] Halle, M. (1985). Speculation about the representation of words in memory. In V. Fromkin (Ed.), *Phonetic Linguistics* (pp.101-114.) New York: Academic Press

[9] Kučera, H., & Francis, W. N. (1967). *Computational analysis of present day American English*. Providence, RI: Brown University Press.

[10] Baayen, R.H., Piepenbrock, R., & Gulikers, L. (1995) The CELEX Lexical Database (Release 2) [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [distributor].

[11] Nusbaum, H. C., Pisoni, D. B., & Davis, C. K. (1984) Sizing up the Hoosier mental lexicon: measuring the familiarity of 20,000 words. *Research on Speech Perception Progress Report, No. 10.* Bloomington: Indiana University, Psychology Department, Speech Research Laboratory.

[12] Cohen, J., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments, & Computers,* 25, 257-271.

[13] Goldrick, M. (2004). Phonological features and phonotactic constraints in speech production. *Journal of Memory and Language,* 51, 586-603.

[14] Goh, W. D., & Pisoni, D. B. (1998). Effects of lexical neighbourhoods on immediate memory span for spoken words: A first report. *Research on spoken language processing Progress Rep. No. 22.* Bloomington, IN: Indiana University, Department of Psychology, Speech Research Laboratory.

[15] Pierrehumbert, J, B. (2002). Word-Specific Phonetics. In C. Gussenhoven and N. Warner (eds.) *Laboratory Phonology VII,* Mouton de Gruyter, Berlin. 101-140.