Journal of Phonetics 39 (2011) 132–142

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/phonetics

Specificity and abstractness of VOT imitation

Kuniko Nielsen*

Oakland University, Linguistics Department, 320 O'Dowd Hall, Rochester, MI 48309-4401, USA

ARTICLE INFO

Article history:
Received 18 February 2008
Received in revised form
23 December 2010
Accepted 28 December 2010

ABSTRACT

The imitation paradigm (Goldinger, 1998) has shown that speakers shift their production phonetically in the direction of the imitated speech, indicating the use of episodic traces in speech perception. Although word-level specificity of imitation has been shown, it is unknown whether imitation also can take place with sub-lexical units. By using a modified imitation paradigm, the current study investigated: (1) the generalizability of phonetic imitation at phoneme and sub-phonemic levels, (2) word-level specificity through acoustic measurements of speech production; and (3) automaticity of phonetic imitation and its sensitivity to linguistic structure. The sub-phonemic feature manipulated in the experiments was VOT on the phoneme /p/. The results revealed that participants produced significantly longer VOTs after being exposed to target speech with extended VOTs. Furthermore, this modeled feature was generalized to new instances of the target phoneme /p/ and the new phoneme /k/, indicating that sub-lexical units are involved in phonetic imitation. The data also revealed that lexical frequency had an effect on the degree of imitation. On the other hand, target speech with reduced VOT was not imitated, indicating that phonetic imitation is phonetically selective.

1. Introduction

Traditional accounts of speech perception assume that the phonological representation is abstract, and that in perception this representation needs to be extracted from variant speech signals (e.g., Halle, 1985). According to this view, certain types of variability in the speech signal represent "noise" which listeners have to filter out. Support for the notion of abstract representations has traditionally been provided by phonological alternations and phonotactic constraints, where categorically expressed (by distinctive features or gestures) phonological environments solely account for the systematic application of the changes or rules.

In the last two decades, this traditional view has been challenged by exemplar-based theories, which do not necessarily assume linguistic representations to be abstract (Tenpenny, 1995). The episodic view of memory assumes that every stimulus leaves a unique trace (or exemplar) in memory (Shacter, Eich, & Tulving, 1978). Each time a new stimulus is presented, all traces are activated according to their similarity to the stimulus. Recognition of the stimulus takes place when the most activated traces come to consciousness. This view was implemented by exemplar-based models (e.g., Hintzman, 1986; Nosofsky, 1986) in which detailed information in the speech signal is preserved as an exemplar. A perceptual category is defined as the set of all the exemplars of the category, and categorization is based on sums of similarity over all exemplars within each category. Other less-extreme, and linguistically more informed, exemplar-based models include Johnson (1997), Pierrehumbert (2001, 2002, 2006), and Coleman (2002). For example, Pierrehumbert's (2002) modular feed-forward model of speech production is a hybrid of a traditional modular view and an exemplar view: In this model, details of allophonic variability as well as speaker information are preserved in memory and systematically associated with words, capturing attested word-specific phonetic changes. In Coleman (2002), word forms are stored as memories of phonetic (psychophysical) experiences, and statistical regularities over the phonetic space are expressed as phonological constraints.

In fact, a growing number of studies have shown that traces of episodic memory are retained and used in speech perception: listeners perceive speech faster and more accurately when it is repeated in the same voice (Goldinger, 1996; Mullennix, Pisoni, & Martin, 1989), intonation contour and fundamental frequency (Church & Schacter, 1994), or speech style (McLennan, Luce, & Charles-Luce, 2003).

Goldinger (1998) revealed that the effect of episodic traces is also present in speech production: in a single-word shadowing task, subjects shift their production in the direction of the model speech compared with their baseline production (spontaneous imitation). His results also confirmed the word-specific advantage predicted by the extreme exemplar model MINERVA 2 (Hintzman, 1986): stronger imitation effects were observed for low-frequency words than for high-frequency words, as well as among

* Tel.: +1 248 370 2175; fax: +1 248 370 3144.
E-mail address: nielsen@oakland.edu

0095-4470/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wocn.2010.12.007

subjects who were exposed to a higher number of repetitions during the listening block. Later, Goldinger (2000) and Goldinger and Azuma (2004) replicated this imitation effect in a non-shadowing paradigm, in which speakers' post-training productions were recorded five days after they listened to the model speech. Shockley, Sabadini, and Fowler (2004) extended Goldinger's work by showing a significant Voice-Onset-Time (VOT) imitation effect in single-word shadowing for voiceless stops with artificially extended VOTs. Further, Pardo (2006) showed that imitation (or phonetic accommodation) can also occur in more socially rich interactions. More recently, Delvaux and Soquet (2007) examined the influence of ambient speech by employing two regiolects of Belgium French as modeled speech, and showed similar patterns of implicit imitation in segment durations and mel-frequency cepstral coefficients.

In addition to spontaneous imitation, the exemplar view also provides an account for the attested plasticity in speech production (e.g., Pardo, 2006; Sancier & Fowler, 1997), use of phonological categories (e.g., Maye, Aslin, & Tanenhaus, 2003; Norris, McQueen, & Cutler, 2003) or allophones (Carlson, German, & Pierrehumbert, unpublished), as well as for the role of frequency of occurrence in phonetic reduction and phonological change (Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Pierrehumbert, 2002; for reviews).

Taken together, it is evident that neither an abstract view nor an extreme exemplar view alone can account for all the findings in the literature. Therefore, the crucial question to be asked is not *whether* phonological representations are abstract or episodic, but rather *how* abstract and/or episodic they are, and what levels of representations are involved in the process of imitation. A growing body of research addresses these questions, and suggests that sub-lexical abstraction takes place in speech perception. Norris et al. (2003) demonstrated that listeners trained on stimuli containing ambiguous /s/-/f/ phonemes subsequently showed an appropriate shift in their /s/-/f/ categorization boundaries (perceptual learning). At the same time, listeners who heard the ambiguous sound in non-words did not shift the categorical boundary, revealing that listeners utilize their lexical knowledge to dynamically tune their phonemic representations. Later, Eisner and McQueen (2005) reported that perceptual learning in the /s/-/f/ continuum is speaker specific and the learning occurs at a phonemic level, because the effect was generalized to "new" speakers only when the critical phoneme actually was produced by the original speaker (but the vowel was produced by new speakers). McQueen, Norris, and Cutler (2006) showed that the effect of perceptual learning was generalized to novel items which contained the target phoneme, suggesting the locus of the adjustment underlying the perceptual learning to be sub-lexical. Further, Kraljic and Samuel (2006) showed that perceptual learning occurs at a sub-phonemic level: after hearing ambiguous /d/-/t/ phonemes during a lexical decision task, listeners' categorical boundary shift was generalized to /b/-/p/ continua. Goldrick (2004) showed that sub-phonemic abstraction also takes place in phonological learning, in which learned phonotactics were generalized from /f/ to /v/.

Although these results provide support for sub-lexical perceptual representation, little is known in terms of how changes in phonological representations affect one's speech production. Given the previous findings on dissociation between perception and production (e.g., near-mergers by Labov, 1994), finding the production counterpart of sub-lexical generalization will provide a stronger argument for the role of sub-lexical levels of representation, and will yield insight into how closely perception and production interact with each other. As mentioned above, Goldinger's (1998) results from spontaneous imitation provided evidence for an episodic view of the mental lexicon. However, his results do not reveal whether sub-lexical units were also influenced, because the post-exposure productions were elicited in the form of shadowing, and thus the listening (=training) and production lists had to be identical. The present study extends the earlier studies of spontaneous imitation by using a non-shadowing task, which allows the listening and production lists to differ and thus "novel" (or, unheard) words to be introduced into the production list. Using this modified experimental paradigm, we aim to examine the generalizability of the imitation effect at two levels of sub-lexical units, namely, phonemic and sub-phonemic representations.

The feature manipulated in this study is VOT, or aspiration ([+spread glottis]), on the modeled phoneme /p/. VOT was chosen due to its attested imitability in shadowing (Shockley et al., 2004) as well as its ease of acoustic manipulation for stimulus construction.[1] The generalizability of VOT imitation was tested at phonemic and sub-phonemic levels by introducing two types of novel stimuli that were not in the target speech: /p/ initial words (which share the same phoneme with the target words) and /k/ initial words (which share the same feature with the target words). If imitation of positive VOT can be generalized to a different word or phoneme, it indicates that listeners code the information in incoming speech at the phoneme and/or feature level. This will further suggest that the unit of phonological representation (whether it is abstract or episodic) responsible for the imitation effect has to be smaller than a lexical word.

In addition to the use of a non-shadowing task, another innovation of the current study is the method of imitation assessment. In Goldinger (1998, 2000) as well as Pardo (2006), degree of phonetic imitation or convergence was measured through AXB perceptual assessments: a different set of listeners heard three versions of the same lexical items, and judged which item produced by the shadowing talker, A or B (pre-task or shadowed token), sounded like a better imitation of the model X. While overall perceptual assessments integrate multiple acoustic phonetic dimensions and thus provide a more configural and holistic (and thus potentially more powerful) measure of imitation itself, it does not provide information about what is being imitated. Not only will acoustic measurements of one phonetic dimension provide a more objective and precise measure of the imitation itself, it will also provide detailed, quantitative information about the variability among speakers and lexical items. For these reasons, the current study assesses the degree of imitation by acoustic measurements of VOT.

We also aim to replicate the word-level specificity effect observed in Goldinger (1998) through acoustic measurement of one phonetic feature (i.e., VOT). As mentioned earlier, the exemplar view predicts a stronger specificity effect for low-frequency words than for high-frequency words. The exemplar view also predicts a stronger specificity for more recently experienced words, and thus we would expect a larger imitation effect for target words (to which subjects are exposed in the experiment) than for novel words (to which they are not exposed). In Pardo (2006), interacting talkers increased their similarity in phonetic repertoire (phonetic convergence) during conversation, and the convergence persisted into a post-task session. Pardo argued that an episodic memory system containing detailed lexical episodes could not have been the basis of the observed pattern of convergence, because the delay (post-task) would increase the influence of long-term memory traces on repeated words (and consequently decrease the strength of episodic traces), predicting reduced phonetic convergence in the post-task items compared with the during-task items. However, phonetic imitation (in her

[1] Note that the imitability of VOT in non-shadowing paradigms is unknown. In fact, a perceptual adaptation study by Cooper (1979) showed that participants decreased their VOT after the extensive exposure to a target syllable with positive VOT (80 ms).

case, phonetic convergence) persisting into a post-task session does not necessarily refute the effect of episodic traces. In fact, Goldinger (2000) and Goldinger and Azuma (2004) showed that the effect of imitation as well as lexical frequency and number of exposures were still present five days after the training sessions. Although the current study involves a non-social setting, it is similar to Pardo's post-task session in terms of the timing (the test recording is made after the listening session ends). If we observe word-level specificity through degree of imitation, it will provide further support for the role of episodic traces in phonetic imitation.

Lastly, it is also our goal to examine the linguistic selectivity of phonetic imitation, by comparing the imitation pattern of two different types of modeled stimuli, namely extended and reduced VOT. Previous studies have shown that phonetic imitation can be modulated by social factors such as the gender of participants (e.g., Namy, Nygaard, & Sauerteig 2002; Pardo, 2006). However, little is known in terms of the effect of linguistic factors on imitation. Linguistically speaking, imitation of reduced VOT is not parallel to imitation of extended VOT: For voiceless stops in English, increasing VOT has no *phonological* consequences, since there is no phonemic category along the direction of increasing VOT. On the other hand, shortening VOT might impair phonemic contrast with the voiced category (e.g., /p/ vs. /b/) which could further introduce lexical/semantic ambiguity (e.g., *pear* vs. *bear*). Contrast preservation has been argued to be an essential part of phonological grammar (Flemming, 2001; Lubowicz, 2003), and thus shortening VOT should be phonologically highly marked. If phonetic imitation is a process which is sensitive to linguistic structure/grammar, we would expect attenuated phonetic imitation for modeled speech with shortened VOT.

In order to address these issues, two experiments were conducted employing a modified version of the word-naming imitation paradigm (Goldinger, 2000), in which the participants first read stimulus items aloud, then listened to target speech, then read aloud the stimulus items once again. Experiment 1 investigates the imitability of target speech with *extended* VOT on the voiceless stop /p/, and its specificity and generalizability at three levels of phonological representations (i.e., word, phoneme, and sub-phonemic feature). Experiment 2 investigates the linguistic selectivity of phonetic imitation by using the modeled listening stimuli with *reduced* VOT.

2. Experiment 1: Imitation of extended VOT

2.1. Method

2.1.1. Stimuli selection

The production list consisted of 150 words: 120 test words and 30 filler words. Among the test words, 100 were words beginning with /p/ (80 *target* words played in the listening phase, further divided into 40 high-frequency words and 40 low-frequency words, plus 20 *novel* words which were not played during the listening phase), and 20 were *novel* words beginning with /k/. All novel words had low frequency. The filler words always had initial sonorants. The listening list was a subset of the production list, and consisted of 120 words, including the 80 *target* words from the production list (i.e., the 40 high-frequency words and 40 low-frequency words beginning with /p/), and 40 filler words that were different from the ones in the production list. Table 1 shows examples of the test words used in the experiment.

The lexical frequency was determined from both Kučera and Francis (1967) and CELEX2 (Baayen, Piepenbrock, & Gulikers, 1995): the thresholds for low- and high-frequency words were below 5 and above 50 (per million) in Kučera & Francis, and below 300 and above 1000 (per 17.9 million) in CELEX2. Mean frequency (per million) for low- and high-frequency words were 3.48 and 150.38 in Kučera & Francis, and 3.96 and 181.76 in CELEX2, respectively. For the test words (i.e., words with initial /p/ and /k/), other lexical or phonological factors known to influence word recognition and/or production, such as phonological neighborhood density (Luce, 1986; Scarborough, 2004; Vitevitch & Luce, 1999; Wright, 1997, 2004), word familiarity (Wright, 1997), syllable length, and stress pattern, were counterbalanced between the two frequency groups. The phonological neighborhood density (DensityB) and word familiarity (Hoosier Mental Lexicon scale, Nusbaum, Pisoni, & Davis, 1984) were obtained from the Washington University in St. Louis Speech and Hearing Lab Neighborhood Database,[2] and were controlled between the two frequency groups for a given syllable length (i.e., monosyllabic, disyllabic, and trisyllabic words). Mean neighborhood density for high frequency and low frequency words were 7 and 7.2, and mean word familiarity for high frequency and low frequency words were 6.95 and 6.78, respectively. All test words had initial stress, while the stress pattern of fillers (i.e., words with initial sonorants) was varied. There were no words with initial onset clusters or initial voiced stops.

Table 1
Examples of test stimuli used in Experiment 1. Participants only listened to Target words (shown in the darker grid) during the listening phase, while they produced all the test words in both baseline recording and post-exposure recording blocks.

Frequency/ stimulus type	Target /p/ (80)	Novel /p/ (20)	Novel /k/ (20)
High	parent, power	-	-
Low	pebble, pirate	pillar, portal	canine, kosher

2.1.2. Stimulus construction

A phonetically trained male American English speaker served as the model speaker, and provided recordings of the 80 target words in the listening list. The speaker was first asked to read each word twice with normal aspiration, and then twice with extra aspiration. These tokens were recorded in a sound booth located in the UCLA Phonetics Laboratory, and were digitized at 22,100 Hz. Later, the subjectively clearer token of each word was chosen (for normal and hyper-aspirated tokens), and the VOTs for the normally produced initial /p/ were measured by the author (mean = 72.46 ms, SD = 12.14 ms). To construct the modeled speech with artificially extended VOT, the burst and aspiration of normally produced tokens (the period covered by VOT) were replaced by those of hyper-aspirated tokens, such that the resulting VOT was extended by exactly 40 ms. For example, suppose VOT of the normally produced "pass" is 68 ms. The original burst and aspiration (=68 ms) were first deleted, and the rest was spliced with the initial 108 ms of the hyper-aspirated version of "pass" to make its VOT 108 ms. This splicing method was chosen, as opposed to extending the most stable part of aspiration as in Shockley et al. (2004), in order to maximally preserve natural formant transitions from the burst to aspiration. In addition, to ensure that VOTs of manipulated stimuli are substantially longer than those of the normally produced tokens, the minimal VOT threshold was set at 100 ms and further lengthening of VOT was conducted for the seven words whose original VOTs were below 60 ms by splicing with longer portions of the hyper-aspirated versions. For example, for a word whose original VOT was 53 ms, the VOT was extended by 47 ms

[2] http://neighborhoodsearch.wustl.edu/Neighborhood/NeighborHome.asp.

(instead of 40 ms) by splicing with the initial 100 ms of the hyper-aspirated version of the word. Overall mean VOT of modeled speech was 113.26 ms, with standard deviation 10.82 ms.

2.1.3. Procedure

The experiment consisted of four blocks: (1) warm-up reading, (2) baseline recording, (3) target exposure (listening), and (4) post-exposure recording. Each session typically lasted for twenty minutes. Participants were tested individually in a sound booth located in the UCLA Phonetics Laboratory, equipped with a PC, a microphone (TELEX M-540), and headphones (SONY MDR-V250). The experimental stimuli were presented using Psyscope 1.2.5 (Cohen, MacWhinney, Flatt, & Provost, 1993).

In the warm-up reading block, the words in the production list were visually presented on a monitor (one at a time, every 2 s) and the participants were instructed to read the words silently without pronouncing them. This warm-up block was added to the word-naming imitation paradigm (e.g., Goldinger, 2000; Goldinger & Azuma, 2004) in order to reduce possible hyper-articulation for first readings of low frequency words in the baseline recording, which was observed in our pilot data. In the baseline recording block, the same words were presented on a monitor again, but this time the participants were instructed to read the words *aloud* into the microphone, providing a baseline recording. The actual wording of the instruction was the following: "Please identify the word you see on the screen by speaking it into the microphone, as naturally as possible." In the target exposure block, the participants were asked to carefully listen to the modeled speech (two repetitions of the words in the listening list) using headphones. There was no additional task during this block. Finally, in the post-exposure recording block (which was identical to the baseline recording block), the participants were instructed to produce the words in the production list for the second time, providing a post-exposure recording. Across the four blocks, the words were presented in random order for each subject. Participants' tokens were digitally recorded into a computer at a sampling rate of 22,100 Hz, and VOTs and whole-word durations were later measured from both waveforms and spectrograms using Praat (Boersma, 2001) by phonetically trained research assistants. Unlike previous studies, there was no perceptual assessment (i.e., AXB testing) of the baseline vs. post-exposure productions.

2.2. Results and discussion

2.2.1. Statistical analysis

Statistical analysis of the data was based on mixed-effects modeling (*cf.* Baayen, 2008) using the lmer function in the lme4 package for R (R Development Core Team, 2008). The response (outcome measure) was the percent increase in duration between the baseline and post-exposure utterances [calculated as 100(post-exposure/baseline − 1)]. The percent increase was calculated separately for VOT and for the rest of the word (=*Rest*), and the two values were treated as repeated measures. This approach makes it possible to determine whether changes in VOT could be due to global changes in speech style as opposed to a shift in distribution of exemplars.

The basic model is a straight-line regression of the percent increase in duration on the logarithm of the Kučera & Francis measure of lexical frequency for the words used in the study. Both the intercept and the slope of the regression line were allowed to depend on Word Type (Target /p/ low, Target /p/ high, Novel /p/, and Novel /k/) and on Word Part (VOT, Rest). In addition, random effects for the intercept (based on both Subjects and Words) and slope (Subjects only) were included. Likelihood ratio tests were used to determine which effects were needed in the model.

The question of how many "degrees of freedom" are available for the traditional t statistics is unresolved and currently quite controversial. However, the data set in this experiment is large enough to make this a relatively minor issue in this case (*cf.* Baayen, Davidson, & Bates, 2008). A relatively conservative value of degree of freedom of 2000 was used in the following analyses (even this conservative approach is virtually indistinguishable from using a normal distribution instead of t). In addition, Markov Chain Monte Carlo (MCMC) based p-values, which were obtained when possible, were quite similar to the t distribution p-values.

2.2.2. Mixed-effects modeling

In order to study the structure of random effects, a complex model with separate regression lines for each of the eight conditions was used. The initial analysis revealed that there was no difference between Target /p/ high and Target /p/ low, and thus a model was developed which combined the two groups and involved twelve distinct parameters (see Table 2). Using this model, the contributions of the random effect terms were assessed. The random intercept effect for Word was found to have no impact [variance estimated as zero, $\chi^2=0$, df$=1$, $p=1$], so this term was dropped from the model. In contrast, the Subject random effects for the intercept [$\chi^2=861.3$, df$=1$, $p<0.0001$] and the slope [$\chi^2=25.4$, df$=1$, $p<0.0001$] were both significant. In fact, a model that included separate Subject effects for the two Word Parts (i.e., VOT and Rest) was used because it provided a substantially better fit to the data than a model with a single random effect.

Although not significant ($p>0.1$), there was an indication that the random Subject effects for the slope and intercept should be treated as correlated. However, using a model with uncorrelated effects provided essentially the same fit, and also made it possible to use MCMC sampling for inference, which produces more trustworthy p-values.

After determining the structure for the random effects, analysis of the parameters of the six regression lines was performed. MCMC based p-values were obtained for the individual parameters, and t-tests were performed to test for differences among them. As can be seen in Table 2, the intercepts for Target /p/ and Novel /p/ were quite similar to each other (for both VOT and Rest), while the intercepts for Novel /k/ were much lower than those for /p/. As for the slopes, it was only Target /p/ which showed negative correlations for both VOT and Rest. Statistical comparison of these parameters will be presented in the following sections.

2.2.3. Phonetic imitation

Table 3 shows a summary of the durational measurements from Experiment 1. The imitation of extended VOT was observed in a non-shadowing paradigm [$t=5.937$, $p<0.001$]. Fig. 1 summarizes the results from Experiment 1, and plots VOT in milliseconds under three types of stimuli. The light bars show the mean VOT of the baseline production, and the dark bars show the post-exposure production. The error bars represent the standard error of the mean. As seen in the figure, post-exposure

Table 2
Summary of regression coefficients for Experiment 1.

Word part	VOT		Rest	
Word type	**Intercept**	**Slope**	**Intercept**	**Slope**
Target /p/	17.83	−0.71	5.6	−0.13
Novel /p/	16.92	+0.85	5.2	+2.06
Novel /k/	8.14	−1.90	3.7	+0.84

Table 3
Summary of Experiment 1 results. The mean, median, and standard deviation of VOT and Rest (=word duration – VOT) are shown (in milliseconds) for each stimulus type.

Word type	Word part	Production type	Mean (ms)	Median (ms)	Std. deviation (ms)
Target /p/ **low**	**VOT**	**Baseline**	65.9	64.6	20.0
		Post-exposure	73.6	72.6	21.9
	Rest	**Baseline**	441.8	434.3	125.5
		Post-exposure	448.9	441.6	130.6
Target /p/ **high**	**VOT**	**Baseline**	66.0	64.1	20.6
		Post-exposure	72.4	70.5	22.0
	Rest	**Baseline**	483.6	480.6	124.2
		Post-exposure	492.4	487.8	127.4
Novel /p/ **low**	**VOT**	**Baseline**	63.0	61.5	19.9
		Post-exposure	69.8	68.5	20.7
	Rest	**Baseline**	458.1	452.5	107.0
		Post-exposure	463.5	450.7	116.5
Novel /k/ **low**	**VOT**	**Baseline**	75.6	73.6	19.0
		Post-exposure	81.2	82.3	19.0
	Rest	**Baseline**	473.1	472.9	105.2
		Post-exposure	479.7	470.9	109.3

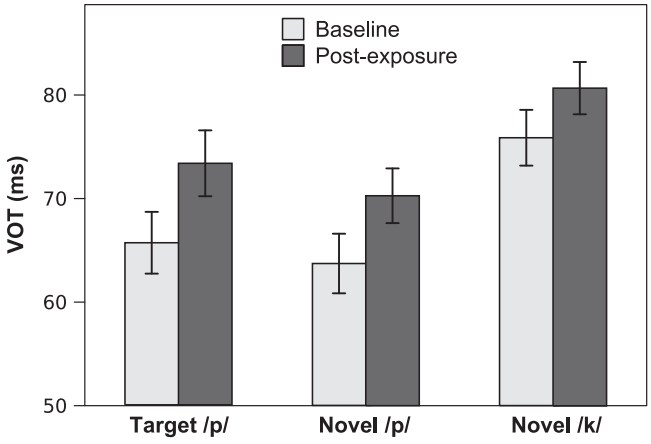

Fig. 1. Generalizability of spontaneous imitation. Mean VOT and standard error of the mean in Experiment 1 (Extended VOT), presented separately for the baseline and post-exposure productions, and for the three types of stimuli. The participants heard only Target /p/ (low and high frequency combined). For each stimulus type, the difference between baseline and post-exposure VOTs was significant.

productions show consistently longer VOTs than baseline productions in all types of stimuli, revealing that the VOT imitation effect is present even when the task involves non-shadowing elicitation-style production. This finding is consistent with previous studies (e.g., Goldinger & Azuma, 2004); several minutes after the participants heard the target speech, they sustained and imitated the modeled speech's detailed surface phonetic information (i.e., extended aspiration), without being instructed to do so.

The increase in VOT durations was significantly larger (17.83% for Target words; see Table 2) than that in Rest (estimated as 5.6%) [$t=3.229$, $p<0.001$], revealing that the speakers increased their VOT proportionally more than they increased the word durations. Given these results, it is unlikely that global aspects of speech are solely responsible for the spontaneous phonetic imitation observed here.

2.2.4. Generalizability

As can be seen in Fig. 1, VOTs of Novel items (both /p/ and /k/), which the participants were not exposed to during the listening block, increased in the post-exposure productions. The change was statistically significant [Novel /p/: $t=4.075$, $p<0.001$; Novel /k/: $t=2.043$, $p<0.03$]. These results reveal that the imitation of extended VOT was generalized at a sub-lexical level. Next, degrees of generalized VOT imitation for Novel /p/ and Novel /k/ were compared in order to determine whether there was a phoneme-specific effect. The mixed-effects model showed a significant difference between the two [$t=-2.185$, $p<0.03$], revealing that the magnitude of generalized imitation was larger for Novel /p/ than for Novel /k/. These findings suggest that the extended VOT in the modeled speech was coded at *both* phoneme *and* feature levels, and subsequently affected the two levels of representations independently.

2.2.5. Word specificity

In order to replicate the effect of word specificity (Goldinger, 1998, 2000) through acoustic measurements of a phonetic feature, the current study controlled both Lexical Frequency and Word Type (Target /p/, Novel /p/, and Novel /k/) as fixed effects variables. Word specificity would mean that low frequency words should show a stronger imitation effect, and so should Target words. To test the effect of lexical frequency, the log frequency from Kučera and Francis (1967) was used in the current analysis (as opposed to CELEX2) as it was the source of lexical frequency analysis in Goldinger (1998, 2000). A significant effect of lexical frequency was found in the mixed-effects modeling [$t=-2.138$, $p<0.04$], replicating the lexical frequency effect found in Goldinger (1998, 2000).

The comparison between Target /p/ and Novel /p/ revealed no significant difference in terms of their VOT intercepts [$t<1$, $p>0.5$] nor slopes [$t<1$, $p>0.5$], despite the fact that their estimated VOT slopes were in different directions (-0.71 vs. $+0.85$, as seen in Table 2). That is, as far as the comparison between Target /p/ and Novel /p/ words was concerned, our data showed no evidence for word-specific patterns of imitation.

2.2.6. Individual variability

In addition to the effects of phonetic imitation, sub-lexical generalization, and word-level specificity, our data revealed a wide range of individual variability in VOT as well as degree of imitation. Fig. 2 illustrates the speaker variability in VOT and

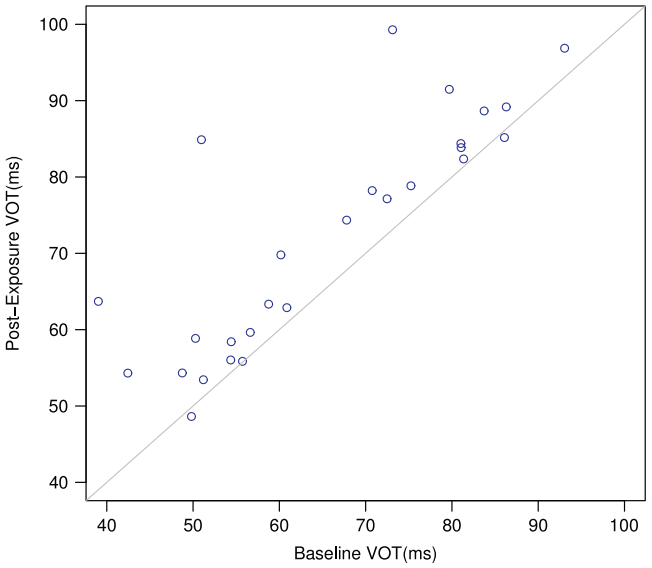

Fig. 2. Individual variability in VOT and phonetic imitation. Mean VOT of individual participants in Experiment 1 (Extended VOT). The baseline VOT is plotted on the x-axis, and post-exposure VOT is plotted on the y-axis ($R^2=0.71$). The diagonal line's slope is one and indicates no imitation (baseline=post-exposure).

degree of imitation in Experiment 1. As can be seen, the mean baseline VOT ranged from 39 to 93 ms. The magnitude of imitation ranged from −3% to +63% of the baseline. Although most participants increased their VOT after they listened to the target speech with extended VOT, the degree of the changes varied considerably from participant to participant. The distribution was normal, with strongly positive skew (mean=1.16, median=1.10, skewness=2.52, kurtosis=6.48).

In addition, our results showed that a model that included separate Subject effects for VOT and Rest provided a substantially better fit to the data than a model with a single random effect. This finding suggests that subject-specific effects for imitating VOT vs. other aspects of the utterance (e.g., whole-word duration) could be based on different mechanisms.

3. Experiment 2: Imitation of reduced VOT

Experiment 2 was designed to investigate the linguistic selectivity of phonetic imitation by using modeled listening stimuli with *reduced* VOT. Unlike extended VOT, imitating reduced VOT in English voiceless stops could introduce phonemic ambiguity (i.e., confusion with the voiced category). If phonetic imitation is a process which is sensitive to phonological structure, we would expect attenuated phonetic imitation for modeled speech with reduced VOT. On the other hand, if phonetic imitation is not linguistically selective, the presence of phonemic contrast should not constrain the effect and thus we would expect similar patterns of imitation for the two types of modeled stimuli.

3.1. Method

3.1.1. Participants

Twenty-five native speakers (12 F and 13 M) of American English with normal hearing served as subjects for this experiment. They were recruited from the UCLA student population and received course credit for participating.

3.1.2. Stimuli selection

The same stimuli as in Experiment 1 (both production and listening lists) were used in Experiment 2.

3.1.3. Stimulus construction

The same original recording of normally produced target words used in Experiment 1 was used to construct listening stimuli for Experiment 2. The most stable part of aspiration was deleted from the original recording, so that the resulting token's VOT was reduced by exactly 40 ms from the original (mean=30.36 ms, SD=8.95 ms). Similar to Experiment 1, the maximal VOT threshold was set at 40 ms to ensure that VOTs of manipulated stimuli were substantially shorter than those of the normally produced tokens, and thus further shortening was conducted for the nineteen words whose original VOTs exceeded 80 ms. Note that of eighty target words in the listening list, there were ten words in which confusion with /b/ was lexically possible (e.g., pall-ball, peck-beck). To ensure that these tokens still sounded like the target words (i.e., initial phoneme /p/ as opposed to /b/), two native English speakers listened to the target words and recorded what they thought they heard. Every word was perceived as a /p/-initial word by both listeners.

3.1.4. Procedure

The same experimental procedure as in Experiment 1 was used, except that only VOT was measured in Experiment 2. Whole-word duration was not measured because the results in Experiment 1 showed that there were only small effects on the rest of the word.

3.2. Results and discussion

3.2.1. Statistical analysis

As in Experiment 1, the statistical analysis was based on mixed-effects modeling. The basic model is a straight-line regression of the percent increase in VOT on the logarithm of the Kučera & Francis measure of lexical frequency. The intercept of the regression line was allowed to depend on Word Type (Target /p/ low, Target /p/ high, Novel /p/, and Novel /k/). Additionally, random effects for the intercept (based on both Subjects and Words) were included.

3.2.2. Mixed-effects modeling

A model with separate regression lines for each condition was used in order to study the structure of random effects. Random intercepts by Word made no contribution to the model [$t<1$, $p>0.1$], and the contribution of random log(KF) slopes by Subject was also not significant [$t<1$, $p>0.1$]. However, random intercepts by Subject made a strong contribution to the model [$\chi^2=715.487$, df=1, $p<0.0001$].

3.2.3. Phonetic imitation

Contrary to the results in Experiment 1, the reduced VOT in the modeled speech was not imitated [$t<1$, $p>0.1$]. Table 4 shows a summary of the results in Experiment 2: the mean, median, standard error, and standard deviation of VOT (ms) are shown by stimulus type. As can be seen, the participants in Experiment 2 did not imitate short VOT in the way the participants in Experiment 1 imitated long VOT.

3.2.4. Word specificity and generalizability

A mixed-effects modeling revealed that none of the fixed effects variables had a significant effect on the degree of imitation [$t<1$, $p>0.1$], except that the change in VOT for Novel /k/ was significantly different from the other Word Types (but none were significantly different from zero) [$\chi^2=4.79$, df=1, $p<0.05$]. In sum, the participants did not significantly change their speech with respect to VOT after being exposed to target speech with reduced VOT, such that word specificity and generalization cannot be evaluated.

Given that the effect of word specificity was tested through the magnitude of imitation in the current study, the presence of clear imitation is a prerequisite. The lack of overall imitation thus could be confounding possible effects of word specificity. For this reason, a post-hoc analysis was performed only for the participants who imitated the reduced VOT to some degree (i.e., whose average VOT in the post-exposure production decreased by more than 5% from their baseline). Among twenty-five participants in

Table 4
Summary of Experiment 2 results. The mean, median, and standard deviation of VOT (ms) are shown for each stimulus type.

Word type	Production type	Mean (ms)	Median (ms)	Std. deviation (ms)
Target /p/ **low**	**Baseline**	66.1	64.4	19.4
	Post-exposure	66.9	66.6	20.8
Target /p/ **high**	**Baseline**	66.8	66.2	20.1
	Post-exposure	66.56	66.1	21.3
Novel /p/ **low**	**Baseline**	65.0	64.4	20.0
	Post-exposure	64.4	64.5	21.4
Novel /k/ **low**	**Baseline**	77.3	76.1	17.2
	Post-exposure	76.2	73.9	17.7

Experiment 2, eight (3 F and 5 M) met this criterion. A mixed-effects modeling revealed no significant main effects of Word Type [$t < 1$, $p > 0.1$] nor Lexical Frequency [$t < 1$, $p > 0.1$]. That is, for the participants who actually imitated the reduced VOT, the degree of imitation did not vary across all types of production stimuli, namely, Target words, Novel words with initial /p/, and Novel words with initial /k/. These results show that when the reduced VOT was imitated, it was generalized at a feature level, while there was no evidence for phoneme-level generalization or word specificity in the data.

3.2.5. Individual variability

Similar to Experiment 1, both VOT values and the degree of imitation varied considerably across participants. Fig. 3 illustrates the individual variability of VOT and the pattern of imitation found in Experiment 2. As seen, the mean VOT value per participant ranged from 44 to 84 ms, while the degree of the changes varied from −19% to +22%. The distribution was normal, with slightly negative skew (mean = 1.03, median = 1.05, skewness = −0.32, kurtosis = −0.36).

3.2.6. Long vs. short VOT (Experiments 1 and 2 comparison)

Fig. 4 summarizes the different patterns of imitation observed in Experiments 1 and 2: As seen, while the baseline VOT values were equivalent, post-exposure VOT values in the two experiments were clearly different. The clear asymmetry of phonetic imitation found in the data, namely the absence of reduced VOT imitation, suggests that spontaneous phonetic imitation is *not* an automatic process which reflects the collection of raw percepts, but rather a process which can be modulated by other factors.

Fig. 5 shows the distribution of VOT in Experiments 1 and 2 (/p/ = initial words only), plotted separately by Baseline vs. Post-Exposure. As can be seen, exposure to the target speech did not change the overall pattern/shape of distribution in either experiment. However, in Experiment 1 (left column), the entire distribution shifted rightward on the *x*-axis (= longer VOT) after target exposure. This indicates that the imitation observed in Experiment 1 is due to the changes in the entire VOT distribution, as opposed to changes in the tails of the distribution.

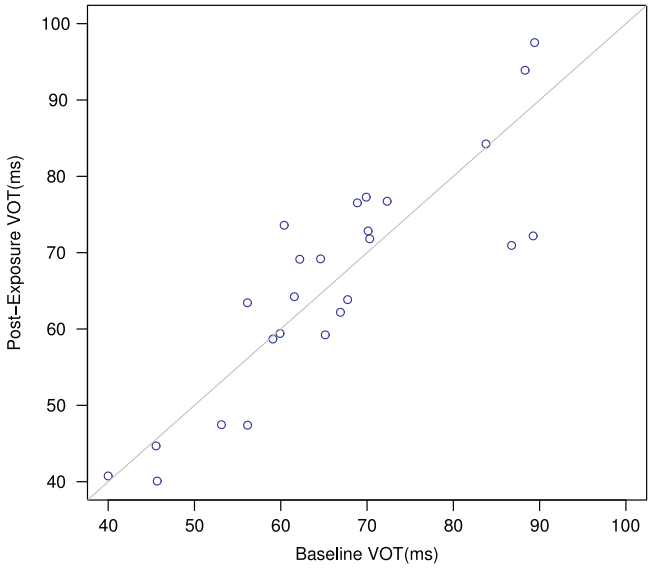

Fig. 3. Individual variability in VOT and phonetic imitation. Mean VOT of individual participants in Experiment 2 (reduced VOT). The baseline VOT is plotted on the *x*-axis, and post-exposure VOT is plotted on the *y*-axis ($R^2 = 0.88$). The diagonal line's slope is one and indicates no imitation.

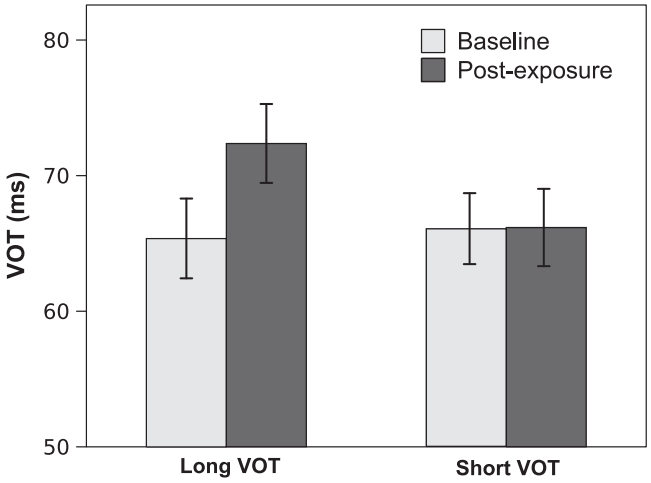

Fig. 4. Comparison of Experiments 1 and 2. Mean VOT and standard error of the mean for words with initial /p/, presented separately for the baseline and post-exposure productions, and for the types of listening stimuli. Artificially extended VOT was imitated, whereas reduced VOT was not.

4. General discussion

4.1. Size/unit of phonological representation

The results in the current study provide evidence in support of word, phoneme, and feature-level representations. Although previous studies have indicated the role of sub-phonemic representation in perceptual and phonotactic learning experiments (Goldrick, 2004; Kraljic & Samuel, 2006), relatively little was known about its production counterpart, except for the previous work on speech errors which shows both sub-phonemic and sub-featural representations (e.g., Fromkin, 1973; Frisch & Wright, 2002; Goldstein, Pouplier, Chen, Saltzman, & Byrd, 2007). Given the previous findings on dissociation between speech perception and production (e.g., Labov, 1994), our finding of sub-lexical and sub-phonemic generalization in phonetic imitation provides additional support for these levels of representations in speech production, as well as their plasticity and flexibility. More importantly, the observed pattern of imitation can only be explained by assuming that (at least) three levels of representations were simultaneously at work.

Note, however, that these sub-phonemic representations do not necessarily correspond to the distinctive features that are widely used in phonology. In fact, the sub-phonemic features which were responsible for the generalization in the data cannot be described in terms of categorical features but have to be more fine-grained, since what we manipulated was the degree of aspiration, not the categorical value of the feature [spread glottis]. In other words, in order to account for the results found in the current study, the sub-phonemic representation (which was imitated and generalized) has to be expressed in a gradient, not categorical manner. This is consistent with our observations of lexical frequency effects, namely that our phonological representations have to include rich phonetic information. Further, our results do not necessarily provide support for acoustic features per se. It is entirely possible that the target of phonetic imitation is articulatory (Browman & Goldstein, 1989; Fowler, Brown, Sabadini, & Weihing, 2003; Shockley et al., 2004), which specifies the coordination between glottal opening and closure. These two theoretically contrastive views (features vs. gestures) cannot be distinguished in the current study.

4.2. Phonological representations: how episodic and abstract?

The current study replicated Goldinger's (1998, 2000) findings on phonetic imitation and its lexical frequency effect through

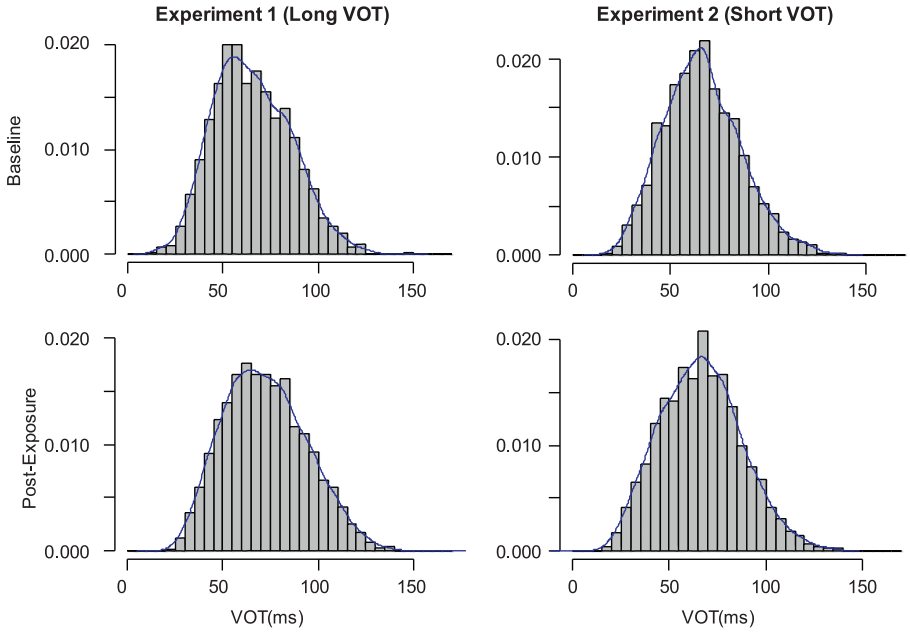

Fig. 5. Distributions of VOT in Experiments 1 (Extended VOT) and 2 (Reduced VOT). Histograms of VOT (/p/=initial words only) distribution in Experiments 1 and 2, presented separately for the baseline and post-exposure productions.

acoustic measurements, providing more precise and quantitative information about the variability among speakers and lexical items. More importantly, however, our data also demonstrated that the imitation was systematically generalized at both phoneme and sub-phonemic levels to words that participants did not listen to during the experiment, and that the effect of generalization was more robust than the observed effect of lexical frequency. These findings are broadly consistent with Carlson et al. (unpublished), who found a robust effect of systematic learning of Glaswegian /r/ (or transfer of an existing allophone to a new phoneme) and small lexical effects.

As for the episodicity of phonological representations, our result was rather inconclusive. Although the finding of phonetic imitation itself does challenge the abstract and invariant view of phonological representation, it is our finding of lexical frequency effects which lends support to the exemplar view, because the concept of a variant representation is compatible with other accounts such as statistical learning or accommodation (*cf.* Giles, Coupland, & Coupland, 1991). At the same time, the observed effects of lexical frequency can also be viewed as a mechanism whereby a *property* of lexical items (e.g., low frequency) influences phonetic imitation (*cf.* Bell et al., 2009). This view would allow for word-level properties to influence phonetic properties without direct associations of phonetic properties with lexical representations, as assumed by the exemplar view. In addition, the other effect of word specificity predicted by the exemplar view (i.e., the effect of target exposure; Target /p/ words would show stronger imitation than Novel /p/ words) was not observed in our data.

As described above, the effect of sub-lexical generalization was more robust in the data compared to the effect of lexical specificity, suggesting that the mental lexicon consists of phoneme- and feature-size representations which are robust and plastic, and also of detailed word-size representations. Note that although the finding of phoneme- and feature-level generalization hints at the "abstractness" of these representations, it neither proves its abstractness nor challenges exemplar theories, so long as sub-phonemic exemplars are assumed. Our experiments were not designed to differentiate between episodic and abstract sub-lexical representations. In fact, if we assume *multiple levels* (or sizes) of exemplars (*cf.* Walsh, Möbius, Wade, & Schütze, 2010), the seemingly large distinction between abstractionist and episodic views will diminish. The key contribution of the current study is exactly this point: Mental representations of sound structures are multi-leveled, and as a consequence they can be both episodic and systematic.

4.3. Selective imitation

Experiment 2 showed that artificially shortened VOT was *not* imitated, revealing that spontaneous phonetic imitation is not an automatic process, as claimed in Gentilucci and Bernardis (2007). This selective nature of phonetic imitation was recently found in Babel (2009) as well: she investigated phonetic imitation of English vowels, and found that not all vowels were imitated as a result of exposure to the model talkers. Her data also showed that social factors such as perceived attractiveness of the modeled talkers and gender of participants modulated imitation, confirming that phonetic imitation is sensitive to both phonetic and social factors.

The exact mechanism which caused the observed asymmetrical pattern of imitation in the current study cannot be determined. The special social setting (i.e., laboratory speech) might have discouraged shortening VOT, as participants typically produce more hyperarticulated speech in laboratory studies (although the lengthening of VOT observed in Experiment 1 was not due to changes in global aspects of speech, as discussed in Section 2.2.1). Nonetheless, the two conditions clearly differ in light of linguistic structure: imitating reduced VOT can introduce phonological ambiguity with the voiced stop, while there are no such consequences in imitating extended VOT. A similar asymmetrical pattern found in the VOT goodness-rating study by Allen and Miller (2001)[3] suggests that native speakers are aware of this difference. Given the small size of change in VOT (+7 ms in

[3] Regardless of conditions, the goodness rating curves are positively skewed (i.e., the rating goes down steeply as the stimuli approach the lower edge along a VOT continuum, while the rating goes down gradually as the stimuli approach the upper edge).

average for /p/ in Experiment 1), however, it is unlikely that the participants in the current study consciously chose to (or not to) imitate based on the goodness of the input. One potential explanation of the observed asymmetry is the depth of processing required in the two conditions. Previous studies have shown that different experimental tasks require various levels of processing. For example, in the speeded categorization task in Miller (2001), response time was relatively long for poor voiceless exemplars with short VOT values (i.e., near the voiced-voiceless category boundary), while it was shorter for poor voiceless exemplars with long VOT values, indicating that categorization time is a function of the distance between the exemplar and the category boundary, not of the perceived category goodness. If the stimuli in Experiment 2 required longer processing times than those in Experiment 1, that would allow more time for the underlying phonemic representation to develop, and possibly override the surface information (i.e., reduced VOT), resulting in no imitation effect.

In addition to the overall asymmetry between the two conditions, large inter-participant variability in the pattern of imitation was observed in both experiments, which is in agreement with previous studies (e.g., Miller, Sanchez, & Rosenblum, 2010; Pardo, 2006). Further research is required to elucidate the mechanism of this inter-speaker variability of imitation. Phonetic imitation appears to be an important component of language change and learning, and understanding when imitation will (or will not) happen will make a contribution to linguistic theories.

4.4. Gesture theories of speech perception

Shockley et al. (2004) and Fowler et al. (2003) argued that gesture theories of speech perception (i.e., the motor theory, Liberman & Mattingly, 1985; the direct realist theory, Fowler, 1994) can account for the phonetic imitation of long VOT they found in single-word shadowing. Unlike traditional acoustic theories, these theories assume listeners' perception of the speech signal to include motor information. Therefore, when participants are asked to shadow a speech signal, their responses are automatically guided by their perception of modeled gestures, thus resulting in spontaneous imitation. Although it is not clear how long the memory of perceived gestures is sustained, our findings of phonetic imitation and its generalization at a sub-phonemic level seem prima facie to be more compatible with both gesture theories than traditional acoustic theories. However, the two gesture theories contrast in terms of their ability to account for the results in Experiment 2, namely the lack of phonetic imitation. The direct realist theory asserts that the objects of perception are distal (such as actual vocal tract movements), predicting phonetic imitation to be automatic. On the other hand, the motor theory asserts that the objects of perception are proximal (such as one's own knowledge of intended gestures) and that speech signal and percepts are independent, and thus phonetic imitation should be non-automatic. Fowler et al. (2003) attributed the discrepancy between modeled and shadowed VOT values to an influence of habitual action on speech production, as opposed to processes that intervene between perception and production. However, habitual action provides an unlikely account for the complete lack of imitation found in Experiment 2 in the current study, because the participants displayed a wide enough range of baseline VOTs to allow them to produce reduced VOT. Although the phenomenon of phonetic imitation suggests a close tie between speech perception and production, our results show that it is not an automatic process, and other factors (such as linguistic contrast, sociostylistic register, and attentional weight) appear to play an important role.

4.5. Implications for models of speech perception and production

Our results suggest that models of speech perception and production should include multiple levels of representations, as well as a capacity to account for other external factors. The levels of representations need to be functionally independent, and potentially vary in their degree of episodicity. Some models of speech perception and production that have recently been proposed appear to be capable of predicting our main findings, namely the imitation and generalization of extended VOT, as well as the absence of reduced VOT imitation.

The hybrid model of speech perception and production proposed by Carlson et al. (unpublished) includes both neo-generative and exemplar components, and is equipped with mechanisms to index various factors such as attentional weight and sociostylistic variations. The neo-generative and exemplar components would readily predict the effects of systematic generalization and word specificity, respectively. There are three levels of processing representations assumed in its perception system: perceptual encoding, phonological parsing, and lexical access. A direct connection between perceptual encoding and lexical access is permitted, which predicts word-specific patterns of phonological changes as well as effects of lexical properties such as frequency. Its phonological parsing represents phoneme-size units, predicting phoneme-level generalization (in addition, incorporating additional levels of representation into the multi-leveled model seems feasible).

According to Grossberg's (1980, 1986, 2003) adaptive resonance theory (ART) of speech perception, speech input activates *items* (which are composed of feature clusters) in working memory, which in turn activate *list chunks* in short-term memory. These chunks then resonate with the input, and the resonance between input and chunk constitutes the percept (Grossberg, 1986). List chunks correspond to possible groupings of items that may vary in size, and could refer to tiered processing levels of representations. To interpret our results within the ART framework, we could assume list chunks to include lexical, phonemic, and featural units. Suppose the extended VOT in speech input was salient to listeners, activating phonemic and featural chunks. The chunks then resonated with the input, constituting the percept which included extended VOT. The percept was subsequently reflected in production, resulting in phonetic imitation and its generalization at sub-lexical levels. On the other hand, suppose the reduced VOT was not salient to listeners. Instead of creating a strong resonance between a featural chunk and the input (with reduced VOT), lexical or phonemic chunks (associated with the default VOT value) were activated and received stronger resonance than other chunks, resulting in no imitation. Even if the reduced VOT was salient to listeners, if it was perceived as a bad exemplar, the "badness" could serve as an inhibitory force on the resonance at the feature level, leading other chunks (associated with the default VOT value) to receive stronger resonance (*cf.* phonemic restoration, Samuel, 1981; Warren, 1984). The resulting percepts thus did not contain the reduced VOT, causing no change in production.

Although it is not a fully developed model, Coleman (2002) provides an account of phonological representations as an alternative to the traditional abstractionist account. This account is unique in that it does not assume any abstract representations such as phonemes. Instead, phonological structure consists of phonetic, statistical and semantic knowledge. In this view, phonological representations are "essentially phonetic, rather than symbolic-phonological" (p. 96). Phonological constituents are statistical regularities over the phonetically rich representations, and contrasts among the categories are represented in two dimensions, a continuous physical scale and continuous probability scale.

The combination of these scales is able to capture sociolinguistic variation, effects of lexical frequency, all the imitation patterns observed in the current study, as well as some phonological processes that require gradient representations. This account appears to be linguistically informed, given that statistical regularities have already been shown to provide a rich source of information in language acquisition (e.g., Jusczyk, Luce, & Charles-Luce, 1994; Saffran, Aslin, & Newport, 1996).

Lastly, the Bayesian network model of multi-level phonetic imitation (Nielsen & Wilson, 2008) provides a detailed, quantitative account of the imitation effect for each of the participants in Experiment 1. By employing Bayesian learning techniques, this model assumes that listeners update their internal phonetically fine-grained multi-leveled probabilistic models in response to perceived target speech (i.e., extended VOT), which subsequently generates imitation. The model is multi-leveled and can correctly predict the level of generalization from /p/ to /k/ and the relatively weak effects of word specificity.

5. Conclusions

The current study demonstrated that perceived fine phonetic details are imitable, indicating that they are retained in listeners' memories, and can subsequently affect their speech production. Multiple levels of phonological representations (i.e., word, phoneme, and feature) were shown to simultaneously contribute to the pattern of phonetic imitation, revealing that words are not purely made up of discrete abstract units, but can be episodic and abstract at the same time. Further, phonetic imitation was shown to be not an automatic process, but rather a selective process which can be modulated by linguistic factors. Taken together, these findings call for a hybrid model of speech perception and production, which incorporates modular sub-phonemic representations, the exemplar-based lexicon, and a mechanism to index both internal (or cognitive) and external factors.

Acknowledgments

This study was supported by an NSF Dissertation Improvement Grant (BCS-0547578, PI: Patricia Keating) and a UCLA Dissertation Year Fellowship. The author would like to thank Robert Kushler for his generous help in statistical analysis, and Patricia Keating, Colin Wilson, and the two anonymous reviewers for their constructive comments.

References

Allen, J. S., & Miller, J. L. (2001). Contextual influences on the internal structure of phonetic categories: A distinction between lexical status and speaking rate. *Perception & Psychophysics*, *63*, 798–810.

Baayen, R. H. (2008). *Analyzing linguistic data*. Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

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].

Babel, M.E. (2009). *Phonetic and social selectivity in speech accommodation*. Doctoral dissertation, University of California, Berkeley.

Bell, A., Brenier, J., Gregory, M., Girand, C., & Jurafsky., D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*, 92–111.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*, 341–345.

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

Carlson, K., German, J., & Pierrehumbert, J.B. (2006). *Reassignment of the flap allophone in rapid dialect adaptation*. Unpublished.

Church, B. A., & 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*, 521–533.

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.

Coleman, J. (2002). Phonetic representations in the mental lexicon. In J. Durand, & B. Laks (Eds.), *Phonetics, phonology and cognition* (pp. 96–130). Oxford: Oxford University Press.

Cooper, W. E. (1979). *Speech perception and production: Studies in selective adaptation*. Norwood, NJ: Ablex Publishing Co.

Delvaux, V., & Soquet, A. (2007). The influence of ambient speech on adult speech productions through unintentional imitation. *Phonetica*, *64*, 145–173.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*, 224–238.

Flemming, E. (2001). Scalar and categorical phenomena in a unified model of phonetics and phonology. *Phonology*, *18*, 7–44.

Fowler, C. A. (1994). *Speech perception: Direct realist theory*. *Encyclopedia of language and linguistics*, Vol. 8. Oxford: Pergamon Press pp. 4199–4203.

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. *Journal of Memory and Languages*, *49*, 396–413.

Frisch, S. A., & Wright, R. (2002). The phonetics of phonological speech errors: An acoustic analysis of slips of the tongue. *Journal of Phonetics*, *30*, 139–162.

Fromkin, V. (Ed.). (1973). *Speech errors as linguistic evidence*. The Hague: Mouton.

Gentilucci, M., & Bernardis, R. (2007). Imitation during phoneme production. *Neuropsychologia*, *45*, 608–615.

Giles, H., Coupland, J., & Coupland, N. (1991). *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge: Cambridge University Press.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1166–1183.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 251–279.

Goldinger, S. D. (2000). The role of perceptual episodes in lexical processing. In Anne Cutler, James M. McQueen, & Rian Zondervan (Eds.), *Proceedings of SWAP spoken word access processes* (pp. 155–159). Nijmegen: Max-Planck Institute for Psycholinguistics.

Goldinger, S. D., & Azuma, T. (2004). Episodic memory reflected in printed word naming. *Psychonomic Bulletin and Review*, *11*, 716–722.

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

Goldstein, L., Pouplier, M., Chen, L., Saltzman, E., & Byrd, D. (2007). Dynamic action units slip in speech production errors. *Cognition*, *103*, 386–412.

Grossberg, S. (1980). How does a brain build a cognitive code?. *Psychological Review*, *87*, 1–51.

Grossberg, S. (1986). The adaptive self-organization of serial order in behavior: Speech, language, and motor control. In E. C. Schwab, & H. C. Nusbaum (Eds.), *Pattern recognition by humans and machines: Vol. 1. Speech perception* (pp. 187–294). New York: Academic Press.

Grossberg, S. (2003). The resonant dynamics of speech perception. *Journal of Phonetics*, *31*, 423–445.

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

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*, 411–428.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson, & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–166). San Diego: Academic Press.

Jusczyk, P. W., Luce, P., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, *33*, 630–645.

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, *13*, 262–268.

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

Labov, W. (1994). *Principles of linguistic change. Vol. 1: Internal factors*. Oxford & Cambridge, MA: Blackwell.

Liberman, A., & Mattingly, I. (1985). The motor theory revised. *Cognition*, *21*, 1–36.

Lubowicz, A. (2003). *Contrast preservation in phonological mappings*. Doctoral dissertation, University of Massachusetts, Amherst. Amherst: GLSA Publications.

Luce, P.A. (1986). *Neighborhoods of words in the mental lexicon*. Doctoral dissertation, Bloomington, IN: Indiana University.

Maye, J., Aslin, R., & Tanenhaus, M. (2003). In search of the weckud wetch: Online adaptation to speaker accent. In *Proceedings of the 16th annual CUNY conference on human sentence processing*, 27–29 March, Cambridge, MA.

McLennan, C. T., Luce, P. A., & Charles-Luce, J. (2003). Representation of lexical form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 539–553.

McQueen, J. M., Norris, D., & Cutler, A. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*, 1113–1126.

Miller, J. L. (2001). Mapping from acoustic signal to phonetic category: Internal structure, context effects and speeded categorization. *Language and Cognitive Processes*, *16*, 683–690.

Miller, R. M., Sanchez, K., & Rosenblum, L. D. (2010). Alignment to visual speech information. *Attention, Perception, & Performance*, *72*, 1614–1625.
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.
Namy, L. L., Nygaard, L. C., & Sauerteig, D. (2002). Gender differences in vocal accommodation: The role of perception. *Journal of Language and Social Psychology*, *21*, 422–432.
Nielsen, K., & Wilson, C. (2008). A hierarchical bayesian model of multi-level phonetic imitation. In: *Proceedings of the 27th west coast conference on formal linguistics*, pp. 335–343.
Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238.
Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.
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.
Pardo, J. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America*, *119*, 2382–2393.
Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee, & P. Hopper (Eds.), *Frequency effects and the emergence of linguistic structure* (pp. 137–157). Amsterdam: John Benjamins.
Pierrehumbert, J. B. (2002). Word-specific phonetics. In C. Gussenhoven, & N. Warner (Eds.), *Laboratory phonology VII* (pp. 101–140). Berlin: Mouton de Gruyter.
Pierrehumbert, J. B. (2006). The next toolkit. *Journal of Phonetics*, *34*, 516–530.
Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.
Samuel, A. G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General*, *110*, 474–494.
Sancier, M., & Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics*, *25*, 421–436.
Scarborough, R.A. (2004). *Coarticulation and the structure of the lexicon*. Doctoral dissertation, UCLA.
Shacter, D., Eich, J., & Tulving, E. (1978). Richard Semon's theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *17*, 721–743.
Shockley, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception & Psychophysics*, *66*(3), 422–429.
Tenpenny, Patricia.L (1995). Abstractionist versus episodic theories of repetition priming and word identification. *Psychonomic Bulletin and Review*, *2*, 339–363.
Vitevich, M., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, *40*, 374–408.
Walsh, M., Möbius, B., Wade, T., & Schütze, H. (2010). Multi-level exemplar theory. *Cognitive Science*, *34*, 537–582.
Warren, R. M. (1984). Perceptual restoration of obliterated sounds. *Psychological Bulletin*, *96*, 371–383.
Washington University in St. Louis Speech and Hearing Lab Neighborhood Database as a web-based implementation of the 20 000-word Hoosier Mental Lexicon. http://neighborhoodsearch.wustl.edu/Neighborhood/NeighborHome.asp.
Wright, R. (1997). Lexical competition and reduction in speech: A preliminary report. *Research on spoken language processing progress report no. 21*. Bloomington, IN: Speech Research Laboratory.
Wright, R. (2004). Factors of lexical competition in vowel articulation. In J. Local, R. Ogden, & R. Temple (Eds.), *Papers in laboratory phonology VI*. Cambridge: Cambridge University Press.