A Hierarchical Bayesian Model of Multi-level Phonetic Imitation

Kuniko Nielsen[1] and Colin Wilson[2]
[1]Oakland University and [2]Johns Hopkins University

1. Introduction[1]

It is well known that listeners adapt, in some sense, to speech that they have recently heard. Words spoken in recently heard voices or accents are recognized more quickly and accurately (Mullennix et al. 1989; Goldinger 1996; Nygaard & Pisoni 1998; Maye et al. 2003; Kraljic and Samuel 2006, 2007; Smith 2007; see Nygaard 2008 for a review). And listeners can become attuned to novel phonetic characteristics of particular speech sounds (Norris et al. 2003), classes of sounds (Maye et al. 2008, Morley 2008), and even individual words (Dahan & Scarborough 2005).

Research on speech production over the past decade has discovered a counterpart to perceptual adaptation. Talkers implicitly imitate the phonetic properties of speech presented in the form of experimental stimuli (Goldinder 1998, Shockley et al. 2004, Nielsen 2007) and of the speech produced by their interlocutors (Pardo 2006, Delvaux & Soquet 2007). The *phonetic imitation effect*, and its theoretical analysis in terms of a model of phonetic knowledge and learning, is the focus of our paper.

We begin by discussing a phonetic imitation experiment that replicates and extends the results of previous work. Participants in the experiment were exposed to speech in which the voice-onset time (VOT) of one word-initial voiceless stop (namely, [p^h]) had been digitally lengthened. When the participants later produced the same words that they had heard as stimuli, and different words beginning with [p^h] that had not heard, their VOTs were longer in comparison to a pre-listening baseline condition. Importantly, the participants also lengthened VOT when producing words that begin with a different voiceless stop, namely [k^h]. This novel result provides evidence against models of speech perception and production that have no representations other than individual sounds or individual words. If such models were correct, participants in this experiment would have had to somehow 'imitate' phonetic manipulations that they had not experienced.

We analyze phonetic imitation in general, and the extension of imitation from [p^h] to [k^h] in particular, with a model that has multiple levels of linguistic representation and a statistically sound mechanism of adapting to experience. For the purposes of this paper, we take the set of levels to include at least word and feature (or gesture) representations. It is the existence of featural/gestural representations that supports generalization across segments: [p^h] and [k^h] share a feature, such as [–voice] or [+spread glottis] ([+s.g.]), that is phonetically interpreted in terms of VOT (among other phonetic properties; Liberman et al. 1958; Lisker & Abramson 1964); informally, then, we expect a manipulation of the VOT of some [p^h]-initial words to be extended to [k^h]-initial words — that is, to lead to phonetic imitation at the level of the feature — as well as to other [p^h]-initial words. Because the level of phonetic imitation is somewhat greater for words that were actually heard with lengthened VOT (see also Goldinger 1998), we characterize the overall pattern as *multi-level* phonetic imitation. All else being equal, we expect that all levels of linguistic representation are engaged in phonetic imitation, and therefore that the size of the imitation effect for a particular word will roughly correlate with its similarity, as defined by the model, to words that have been recently heard.

[1] We would like to thank Adam Albright, Lisa Davidson, Diamandis Gafos, Bruce Hayes, Pat Keating, Elliott Moreton, Paul Smolensky, Kie Zuraw, members of the NYU Phonetics Lab, and the participants of WCCFL 27 for comments and questions on the work reported here. Special thanks to Cheryl Hipp for her expert assistance in carrying out the phonetic analysis reported in Section 2. The authors are listed in alphabetical order.

© 2008 Kuniko Nielsen and Colin Wilson. Proceedings of the 27th West Coast Conference on Formal Linguistics, ed. Natasha Abner and Jason Bishop, 335-343. Somerville, MA: Cascadilla Proceedings Project.

Formalizing our intuitive understanding of phonetic imitation requires a model in which statistical distributions are updated based on perceptual experience, and then applied generatively to yield predictions about performance. The cornerstone of our model is Bayes' Theorem, which is a rational means of estimating distributions from multiple sources of evidence and prior knowledge (MacKay 2003; Chater et al. 2006). The model is structured *hierarchically*: the phonetic distributions of individual speakers (or perhaps episodes) are nested within a population distribution; the population is itself nested within a universal 'superpopulation' that embodies language learners' a-priori knowledge of phonetic distributions. This Bayesian hierarchy, which runs orthogonal to the levels of representation discussed above, allows the model to account for listeners' knowledge of and adaptation to specific speakers (Mullennix et al. 1989, Nygaard and Pisoni 1998, Smith 2007) and, crucially, provides a principled mechanism for generalizing from partial experience of a new speaker. When exposed to the same experimental stimuli, the model predicts generalization of VOT lengthening at the feature- and word- level like that observed in human participants.

The rest of this paper is organized as follows. Section 2 describes the phonetic imitation experiment outlined above in somewhat more detail.[2] Section 3 develops and applies the hierarchical Bayesian model of multi-level phonetic imitation. Because the model is a straightforward application of standard techniques in Bayesian analysis (Gelman et al. 2004, Gelman and Hill 2007; Bishop 2006), we focus on the aspects of the model that are specific to our proposal: multiple levels of linguistic representation, and hierarchical arrangement of speakers within the population and the population within the superpopulation. In Section 4, we conclude the paper with a summary of the model and a brief discussion of directions for improving and expanding our proposal.

2. Multiple levels of representation and phonetic imitation

The experiment reported in this section was motivated by two main goals. The first was to replicate the phonetic imitation effect with a paradigm that does not involve shadowing or immediate repetition (see also Goldinger 2000; cf. Goldinger 1998, Shockley et al. 2004). The second, more important goal was to test for multi-level imitation. To this end, we compare the degree of imitation in words that had been heard with lengthened VOT (*heard* words), all of which were [p^h]-initial, to the degree of imitation in words that had not been presented as auditory stimuli (*unheard* words). The set of unheard words included both [p^h]-initial and [k^h]-initial items. If imitation were restricted to word-level representations, then only the heard words should be affected by the experimental manipulation. On the other hand, if (at least) word-level and feature-level representations contribute to phonetic imitation, then we would expect all of the words to exhibit the effect, with the possibility that the degree of imitation would be greater for the heard [p^h]-initial words than for the unheard [p^h]-initial words, and for [p^h]-initial words than for [k^h]-initial words in general.

2.1. Method

Participants. Twenty-seven monolingual native speakers of English, recruited from the UCLA undergraduate population, participated in the experiment for course credit. All participants reported normal hearing.

Materials and design. 120 real English words were used as materials in the experiment: 100 [p^h]-initial words and 20 [k^h]-initial words. All of the words had initial stress, and the initial consonant was prevocalic in all cases. 80 of the [p^h]-initial words, 40 low frequency and 40 high frequency, were selected for inclusion in a listening phase.[3] These are referred to as the *heard* items. The remaining 20

[2] This and other experiments are presented more fully in Nielsen (2008). See also Nielsen (2005, 2006) for earlier presentations of some of the same material.

[3] The thresholds for low- and high- frequency items in the experiment were < 5 and > 50, respectively, in the Brown corpus (Kučera and Francis 1967) and < 300 and > 1000 in CELEX2 (Baayen et al. 1995). Phonological neighborhood densities were calculated on-line with the St. Louis Speech and Hearing Lab Neighborhood Database (http://neighborhoodsearch.wustl.edu/Neighborhood/NeighborHome.asp), and word familiarity was assessed with the 7-point Hoosier Mental Lexicon scale (Nusbaum et al. 1984).

[p^{h}]-initial words and all 20 [k^{h}]-initial words were low frequency; they are referred to as the *unheard* items. Neighborhood density, familiarity, and word-length were also controlled across the stimulus set. (The complete stimulus set also included sonorant-initial filler words; see Nielsen 2008.)

The 80 heard items were recorded by a phonetically-trained male speaker of American English. The speaker was asked to produce the words first normally, and then with extra aspiration. Aspiration from the latter productions were digitally spliced into the former, approximately 10 to 20ms before the onset of voicing of the following vowel, in order to lengthen the VOT of the initial consonant by 40ms. If initial splicing resulted in a VOT of less than a predetermined threshold (100ms), further lengthening was performed until the threshold was reached. The resulting tokens had a mean VOT of 113ms, sd = 10.82. (By comparison, the unedited normal productions had a mean VOT of 72ms, sd = 12.14). The approximately 40ms of additional VOT in the edited recordings was expected to induce imitative VOT lengthening in the participants' productions.

The procedure for each experimental participant was a modified version of the word-naming imitation paradigm (Goldinger 2000). First, each of the 120 stimulus words was presented on a computer screen, and the participant was asked to read each one silently. Second, the same word list was presented visually again and the participant was asked to read each word aloud. These *baseline productions* were recorded, and the word-initial stop VOTs were measured by the first author and a phonetically-trained research assistant. Third, the participant was instructed to listen carefully to two repetitions of the VOT-lengthened versions of the 80 [p^{h}]-initial heard items. This listening phase of the experiment did not involve any visual stimulus or other task. Fourth and finally, the participants again read aloud the 120 stimulus items as they were presented on the computer screen. These *test productions* were elicited and analyzed in exactly the same way as the baseline productions. The statistical analysis reported below compares the baseline and test productions for each participant.

2.2. Results and discussion

The table in (1) gives the mean VOTs across participants in the baseline and test productions for each type of stimulus.

(1) VOT means (and standard errors) in the baseline and test productions

Heard	baseline	test	Unheard	baseline	test
[p^{h}] Low freq.	65.73	73.41	[p^{h}] Low freq.	63.73	70.27
(40 items)	(2.98)	(3.18)	(20 items)	(2.88)	(2.64)
[p^{h}] High freq.	65.76	72.09	[k^{h}] Low freq.	75.87	80.66
(40 items)	(3.05)	(3.18)	(20 items)	(2.69)	(2.52)

The table shows small but consistent phonetic imitation effect (difference between test and baseline VOTs) of approximately +5 to +7ms across the stimulus categories. A two-tailed paired *t*-test comparing baseline and test productions reveals that the effect is significant ($t(26) = 16.53$, $p < .01$). In contrast, a comparison of whole-word durations of the baseline and test productions for a randomly selected group of participants (N=8) showed no significant difference ($t(7) < 1$); note that this subgroup did show the imitation effect that was observed in the entire set of participants ($t(7) = 4.67$, $p < .01$). The absence of a change in whole-word duration, in combination with a significant change in VOT, suggests that the phonetic imitation was keyed to the particular phonetic manipulation in the auditory stimuli, rather than being a more global modification of speech style or rate.

Repeated measures ANOVAs with production type (baseline vs. test) crossed with other factors (gender: male vs. female, segment: [p^{h}] vs. [k^{h}], and frequency: high vs. low) showed no significant interactions.[4] The lack of an interaction with segment, despite the fact that only [p^{h}]-initial words were

[4] As expected from previous research, there were main effects of the between-participant factor of gender ($F(1,1) = 12.60$, $p < .01$) and the within-participant factor of segment ($F(1,25) = 271.15$, $p < .01$). It has been reported that females tend to have longer VOT values for aspirated stop consonants than males (Swartz 1992; Whiteside et al. 2004), though this difference may be at least partially due to independent differences in rate of speech (Allen et al. 2004). As for the main effect of segment, it is well known that aspirated velar stops such as [k^{h}] tend to have

heard with elevated VOTs, supports the hypothesis that phonetic imitation engages a level of representation lower than that of individual sounds and is consistent with previous findings by Kraljic and Samuel (2006) on perceptual adaptation. This finding motivates the existence of feature-level representations, and is consistent with the non-existence of segment-level representations, in the model presented in section 3.

We were surprised at the lack of an interaction with lexical frequency, given Goldinger's (1998, 2000) findings of word specificity in phonetic imitation, and therefore examined this issue further. It is possible that word-level representations contribute to phonetic imitation, but that this contribution is smaller or less consistent across speakers than the contributions of lower representational levels. To explore this issue, we identified a subgroup of participants (N=15) whose test VOTs were at least 5% greater than their baseline VOTs and tested for an interaction between production type and lexical frequency in this subgroup. A repeated measures ANOVA with production type and frequency as within-participant factors showed significant main effects of both factors (production type: $F(1,13) = 16.20$, $p < .01$; frequency: $F(1,13) = 7.58$, $p < .05$) as well as a significant interaction ($F(1,13) = 5.78$, $p < .05$). Within this subgroup, low frequency lexical items exhibit a larger phonetic imitation effect (approximately +14ms) than high frequency items (approximately +10ms). (The interaction between production type and segment remained non-significant in this subgroup.)

To summarize, the results of this experiment support the claim that phonetic imitation is not restricted to words, or even segments, that have been experienced with a particular phonetic property. Imitation was generalized to unheard words, including those beginning with a segment ([k^h]) that had never occurred initially in the listening phase of the experiment but which is identical on the relevant feature to the segment that had been heard with exaggerated VOT ([p^h]). The results are also compatible with the claim that phonetic imitation is stronger for lower-frequency words than for higher-frequency words, though the contribution of word-level representations to the effect appears to be weak or inconsistent across speakers for reasons that we do not fully understand. The formal model developed in the next section aims to account for the multi-level structure of these findings.

3. Hierarchical Bayesian modeling

The starting point for our model of phonetic imitation is the hypothesis that adaptation in perception and imitation in production reflect a common underlying learning mechanism: namely, the ability to estimate the values of speaker-specific phonetic parameters. The set of parameters, constraints on their values, and their method of interaction determine the space of possible phonetic distributions that a listener can represent and learn. For example, if the parameters made no reference to individual lexical items, or to lexical properties such as frequency, then adaptation and imitation modulated by purely lexical factors would be impossible. Contrariwise, inclusion of parameters that are sensitive to a particular level of representation makes that level a possible locus of adaptation and imitation. We develop our model in the next three subsections, beginning with assumptions about the representational space, then turning to issues of statistical learning, and finally describing how the model can be used to generate predictions about phonetic imitation and comparing the predictions to the experimental results reported in section 2.

3.1. Parametric model of speakers and the speech population

We assume that a listener's internal phonetic model of a speaker *s(i)* takes the form of a statistical generative grammar *g(i)* of *s(i)*'s speech. Focusing on the parameters that are of relevance for modeling adaptation to and imitation of VOT in initial voiceless stops, we specifically assume that *g(i)* includes one parameter *sg(i)* that represents the basic VOT (i.e., the basic expression of [+spread glottis]) for initial voiceless stops produced by this speaker, one parameter *dor(i)* that represents the (positive) deviation from the basic VOT for the dorsal segment [k^h], and a set of parameters {w*(i,j)*}

longer VOTs than aspirated labial stops such as [p^h], both cross-linguistically (Cho & Ladefoged 1999) and in English in particular (Lisker and Abramson 1964; Zue 1980). There was no main effect of the within-participant factor of lexical frequency ($F(1,25) = 2.19$, $p > .01$); see the text for further discussion of this factor.

representing the (positive or negative) deviations in VOT that are attributed to individual lexical items, where *j* is an index over words. We have found this set of parameters to be sufficient for modeling our experimental results, but do not claim that it is necessary; for example, the item-specific parameters could potentially be replaced by parameters that are sensitive to phonetic or phonological properties (e.g., the following vowel; see Port and Rotunno 1979) or to lexical properties other than lexical identity (e.g., lexical frequency or neighborhood density; see Wright 2004; Scarborough 2003, 2004; Munson and Solomon 2004; Baese-Berk and Goldrick, in press). The formal system presented below could also readily accommodate parameters that are sensitive to gender and other variables of sociolinguistic relevance, as well as effects of speech rate and style.

We assume that listeners model the speaker-specific, word-initial VOT distribution for each word *j* as Gaussian (normal) with mean equal to a linear combination of the parameters and a standard deviation that may also be specific to the speaker, as shown in equation (2). Here 'VOT(*i,j*)' denotes a random variable over VOTs (in ms) for word *j* as produced by speaker *i*, '~' means 'distributed as', '$N(x;y)$' is the normal distribution with mean *x* and variance *y*, and '$\mathrm{I}_{\mathrm{sg}}(j)$' is an indicator function that takes on value 1 if word *j* begins with a [+spread glottis] stop, 0 otherwise. Similarly, '$\mathrm{I}_{\mathrm{dor}}(j)$' is an indicator function that takes on the value 1 if word *j* begins with [k^{h}], 0 otherwise.

(2) Listener model of speaker-specific VOT distribution

$$\mathrm{VOT}(i,j) \sim N(\mathrm{I}_{\mathrm{sg}}(j)\cdot\mathrm{sg}(i) + \mathrm{I}_{\mathrm{dor}}(j)\cdot\mathrm{dor}(i) + \mathrm{w}(i,j);\ \sigma(i)^2)$$

The assumption that VOT distributions are normally distributed was made primarily for computational convenience. At least population VOT distributions appear to be positively skewed (Allen and Miller 2001) and if desired this could be enforced by replacing the normal distribution with a lognormal, gamma, or other nonsymmetric continuous distribution (see, for example, Casella and Berger 2002).

The speaker-specific model (2) immediately raises a learning problem. What should the listener's internal model of speaker *s(i)* predict as probable VOT values for an initial stop in a word that the listener has never heard produced by *s(i)* (say, a nonword that occurs only in the context of an experiment)? This general problem becomes even sharper in the particular context of the experiment reported in section 2. The participants in that experiment heard only a subset of the [p^{h}]-initial words, and none of the [k^{h}]-initial words, that they were asked to produce. Yet to a good first approximation they showed the same level of 'imitation' for all words. How did the participants estimate the *dor** parameter of the digitally manipulated speaker *s**, who they had never heard say [k^{h}] in initial position? How did they estimate the *w(j)** parameters for words that had not been presented during the listening phase of the experiment?

The general answer that we propose for questions such as these is that listeners have a population grammar of VOT distribution in addition to speaker-specific grammars, and that they are able to use the population grammar to make inferences about individual speakers in the absence of positive evidence. The population grammar has the same types of parameters as the grammars for individual speakers, but they are interpreted quite differently. If we think of the parameters of a speaker-specific grammar as a random vector (i.e., vector of random variables) $g(i) = <\mathrm{sg}(i), \mathrm{dor}(i), \mathrm{w}(i,1), \mathrm{w}(i,2), \ldots>$, then the vector of population parameters $G = <\mathrm{sg}, \mathrm{dor}, \mathrm{w}(1), \mathrm{w}(2), \ldots, \mathrm{w}(120)>$ gives the mean or expected value of *g(i)*. Just as the VOT of a word uttered by speaker *i* is modeled as a random draw from the distribution of *g(i)*, the grammar *g(i)* is modeled as a random draw from the population grammar. We assume that the distribution is multivariate normal with mean *G* and covariance Σ.

(3) Listener model of population VOT distribution

$$g(i) \sim N(\mathrm{G}; \Sigma), \text{ where } g(i) = <\mathrm{sg}(i), \mathrm{dor}(i), \mathrm{w}(i,1), \mathrm{w}(i,2), \ldots> \text{ and } G = <\mathrm{sg}, \mathrm{dor}, \mathrm{w}(1), \mathrm{w}(2), \ldots>$$

The listener is now not at a total loss in estimating speaker-specific parameters for which direct evidence is lacking. He or she can use what is known about speakers in general, as represented in the population grammar, to 'fill in' plausible values. At first sight, this solution to the learning problem may seem circular or paradoxical: after all, how does the listener estimate the population grammar *G* except through experience with the productions of individual speakers? There is a certain unavoidable circularity here, but it does not lead to paradox or intractability, as we show in the next subsection.

3.2. Hierarchical Bayesian learning of phonetic parameters

Equations (2) and (3) constitute two levels of a hierarchical model of VOT: VOT tokens are generated from speaker-specific distributions; the parameters of speaker-specific distributions are generated from the population distribution. We assume that there is another distribution, the *superopulation* distribution *UG*, from which the population parameters are generated. As its name suggests, this hierarchically highest distribution encodes universal (language-independent) aspects of VOT patterning. For example, if there is statistical structure in the mean VOT values of aspirated stops across languages, as suggested by the survey of Cho and Ladefoged (1999), this could be encoded in the superpopulation distribution from which population *sg* values are drawn. Similarly, cross-linguistic generalizations about the extent to which dorsal place of articulation increases VOT (see again Cho and Ladefoged 1999) could be encoded in the superpopulation distribution of population *dor* values. We also expect the correct superpopulation distribution over word-specific population parameters to have a smaller mean and variance than those for phonetic parameters such as *sg* and *dor*, since word-level VOT differences within a language seem substantially smaller than VOT differences across languages. Because our interest here was in modeling VOT distributions in a particular language and experiment, not in understanding the constrained variation of VOT distributions across languages, the simulations reported here employ a provisional *UG*: each population parameter was assumed to be drawn independently from a normal distribution with fixed mean of 0 and variance of 1.

Estimating the parameters of equation (2) for each speaker *s(i)* and the parameters of equation (3) for the entire population of speakers, given a sample of VOTs for each speaker and the fixed superpopulation, is a standard application of hierarchical Bayesian inference (Gelman et al. 2004, Gelman 2007). Because the learning problem is 'circular' — the speaker-specific and population parameters are interdependent and must be estimated jointly — there is no closed-form expression for the parameter values. However, an iterative procedure can be employed to identify values of all of the parameters that make the VOT productions most likely given the fixed *UG*. Informally (consult the references just cited for details), the procedure is as follows. In a first step, we fix the values of the population parameters *G* (which are initially set according to some guessing procedure) and perform Bayesian inference of each batch of speaker-specific parameters *g(i)* using *G* as a prior distribution and the productions from speaker *i* as data. The *g(i)*'s can be estimated independently and in parallel. The second step is to now treat the estimated speaker-specific parameters as data and perform the same type of Bayesian inference of the population parameters given the superpopulation prior *UG*. Each of the steps has a closed-form solution that is easy to compute. The two steps are alternated until a convergence criterion indicates that the population values have settled near their optimal values: those that maximize the probability of the data while minimizing departures from the fixed superpopulation prior.[5]

In order to concretely instantiate the model and simulate learning, the 120 baseline VOT values for each of the 27 participants in the experiment of section 2 were taken as data. After a short learning period, the model arrived at parameter values that accurately encoded the VOT distribution of each speaker (all correlations between baseline productions and speaker-specific model predictions had $r^2 \geq .98$) and for the population as a whole. For example, the mean baseline VOT for [p^h] across all participants and words was 65.73ms with a large standard deviation (15.25). The model's estimate of the population value of the *sg* parameter matched this closely at 62.87. Similarly, across all participants the difference between the mean baseline VOT for [k^h] and that for [p^h] was +10.25ms. The model's estimate of the population value of the *dor* parameter approximates this as well at 7.79. (The fact that both parameters are lower than one might expect given the mean data reflects a general property of Bayesian learning with priors that favor 0 values, as our provisional *UG* does.)

Clearly, exposure to 27×120 baseline VOTs vastly underestimates the phonetic experience that the participants had prior to the experience. However, we believe that the learning data is large and

[5] Iterative estimation can be performed simultaneously for the means and (co)variances of the distributions in the model, at some computational cost. The simulations reported here reflect estimation of the parameters that determine the means only; all of the variance parameters of the population and speaker-specific distributions were fixed at a large value (10^2). Learning (co)variances is a focus of our on-going development of the model.

realistic enough to reveal general properties of the hierarchical model. The excellent correlation between each participant's baseline productions and the learned model's predictions for that speaker shows an ability to perform perceptual adaptation. Scaling up to many additional speakers (and hence inclusion of many additional batches of speaker-specific phonetic parameters) should not be problematic, since the parameters for all speakers are learned independently and in parallel. The most relevant test of the model, which we turn to next, is whether its knowledge of the speech population supports empirically valid inferences given impoverished data from a new speaker.

3.3. Predictions about imitation

After the model had learned from the baseline productions, we added a new set of speaker-specific parameters (initially set to random values) that were then learned from the same VOT-lengthened stimuli heard by the participants in the listening phase of our experiment. The digitally manipulated speaker is referred to as s^* and the estimated parameters for that speaker as $g^* = <sg^*, dor^*, w(1)^*, \ldots, w(120)^*>$. The population parameters were held fixed during this phase of learning, reflecting the assumption that the participants' internal models of the population evolve too slowly to be substantially changed within the time of a short experiment. (The assumption that brief experiments can affect speaker-specific parameters, perhaps independently of population parameters, is supported by perceptual adaptation to multiple speakers; see Kraljic and Samuel 2007.)

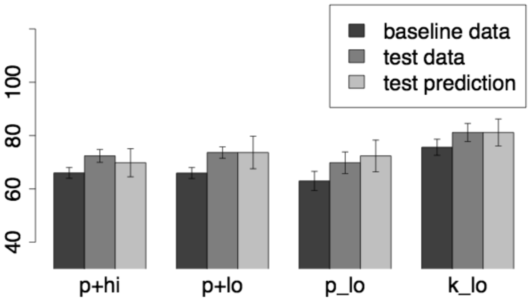
With respect to the 80 heard words, for which positive evidence was provided about s^*, the model showed perceptual adaptation to the new speaker (the correlation between the VOTs of s^* and the model's prediction had $r^2 = .95$ after 20 learning steps). However, this finding alone does not demonstrate an ability to generalize from those words, all of which began with $[p^h]$, to other $[p^h]$-initial words and to $[k^h]$-initial words. Notice in particular that the model had no direct evidence concerning the proper value of the *dor* and several word-specific parameters for s^*.

In order to compare the generalization performance of the model with the imitation results of experiment, we generated predicted test VOT values for each participant $s(i)$ by blending together the speaker-specific models $g(i)$ and g^* with a single mixture parameter $\alpha(i)$ ($0 \leq \alpha(i) \leq 1$) as in (4).

(4) Mixture equation for phonetic imitation

$$\begin{aligned}\mathrm{VOT}(i,j) \text{ at test} \sim\ & \alpha(i) \cdot N(\mathrm{I}_{\mathrm{sg}}(j)\cdot \mathrm{sg}(i) + \mathrm{I}_{\mathrm{dor}}(j)\cdot \mathrm{dor}(i) + \mathrm{w}(i,j);\ \sigma(i)^2) \\ & + [1-\alpha(i)] \cdot N(\mathrm{I}_{\mathrm{sg}}(j)\cdot \mathrm{sg}^* + \mathrm{I}_{\mathrm{dor}}(j)\cdot \mathrm{dor}^* + \mathrm{w}(i)^*;\ (\sigma^*)^2)\end{aligned}$$

The mixture parameter $\alpha(i)$, which embodies participant $s(i)$'s propensity to imitate the experimental stimuli, was fit to the test productions of that participant. Thus there were 27 free parameters, far fewer than the 27×120 test data points. The figure below shows the observed and predicted levels of imitation, with baseline values included for reference, averaged across all participants and binned into four categories: high-frequency heard $[p^h]$-initial words ('p+hi'), low-frequency heard $[p^h]$-initial words ('p+lo'); low-frequency unheard $[p^h]$-initial words ('p_lo'); and low-frequency unheard $[k^h]$-initial words ('k_lo'). Appropriate levels of generalization to the unheard words and segment are evident. (Similar results, not shown here for reasons of space, hold for the individual participants.)



4. Conclusion

In this paper, we have proposed a hierarchically organized model of phonetic adaptation and imitation that employs multi-level linguistic representations and standard principles and techniques of Bayesian learning. This model has a general ability to adapt to experience within the limits of its representational space. When combined with a simple mixture conception of imitative production (equation (4)), the model is able to use what it has learned about individual speakers and the broader speech population to generate empirically valid levels of imitation. In particular, it solves the problem of learning about the phonetic distribution of a new speaker from limited evidence, and thereby correctly generalizes phonetic imitation to unheard words and sounds.

This model, which we have found to be more empirically successful than alternatives that lack either its multi-level representations or its hierarchical Bayesian organization, can be developed and extended in several directions. We have already indicated the open-endedness of the formalism: additional phonetic, phonological, lexical, and sociolinguistic variables could be incorporated into the model without changing its modes of computation and learning. The development we find most pressing, and which has not been addressed in the simulations reported here, is that of estimating the covariances among phonetic properties within a given speaker and across the population.

References

Allen, J. Sean, and Joanne Miller. 2001. Contextual influences on the internal structure of phonetic categories: A distinction between lexical status and speaking rate. *Perception & Psychophysics, 63.5*, 798-810.

Allen, J. Sean, Joanne Miller, and David DeSteno. 2003. Individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America, 113.1*, 544-552.

Baayen, R. Harald, Richard Piepenbrock, and Leon Gulikers. 1995. The CELEX Lexical Database (Release 2) [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [distributor].

Baese-Berk, Melissa, and Matthew Goldrick. In press. Mechanisms of interaction in speech production. *Language and Cognitive Processes*.

Bishop, Christopher. 2006. *Pattern Recognition and Machine Learning*. New York: Springer.

Casella, George, and Roger Berger. 2002. *Statistical Inference*, 2nd edition. Pacific Grove, CA: Duxbury.

Chater, Nick, Joshua Tenenbaum, and Alan Yuille. 2006. Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences, 10.7*, 287-291.

Cho, Taehong, and Peter Ladefoged. 1999. Variation and universals in VOT: evidence from 18 languages. *Journal of Phonetics, 27*, 207-229.

Dahan, Delphine, and Rebecca Scarborough. 2005. Speaker specificity in speech perception: the importance of what is and is not in the signal. *Journal of the Acoustical Society of America, 118.3*, 2034.

Delvaux, Veronique and Alain Soquet. 2007. The Influence of ambient speech on adult speech productions through unintentional imitation. *Phonetica, 64*, 145-173.

Gelman, Andrew, John Carlin, Hal Stern, and Donald Rubin. 2004. *Bayesian Data Analysis*, 2nd ed. Boca Raton: Chapman & Hall/CRC.

Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Goldinger, Stephen. 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, Stephen. 1998. Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*, 251-279.

Goldinger, Stephen. 2000. The role of perceptual episodes in lexical processing. In *Spoken Word Access Processes (SWAP)*-2000, 155-158.

Kraljic, Tanya and Arthur Samuel. 2006. How general is perceptual learning for speech? *Psychonomic Bulletin and Review 13*, 262-268.

Kraljic, Tanya and Arthur Samuel. 2007. Perceptual adjustments to multiple speakers. *Journal of Memory and Language 56*, 1-15.

Kučera, Henry, and W. Nelson Francis. 1967. *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press.

Liberman, Alvin, Pierre Delattre, and Franklin Cooper. 1958. Some cues for the distinction between voiced and voiceless stops in initial position. *Language and Speech, 1*, 153-167.

Lisker, Leigh and Arthur Abramson. 1964. Cross-language study of voicing in initial stops: acoustical measurements. *Word, 20*, 384-422.

MacKay, David. 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press.

Maye, Jessica, Richard Aslin, and Michael Tanenhaus. 2003. In search of the weckud wetch: Online adaptation to speaker accent. In *Proceedings of the 16th Annual CUNY Conference on Human Sentence Processing*, March 27–29, Cambridge, MA.

May, Jessica, Daniel Weiss, and Richard Aslin. 2008. Statistical phonetic learning in infants: facilitation and feature generalization. *Developmental Science, 11:1*, 122-134.

Morley, Rebecca. 2008. Rapid learning of morphologically conditioned phonetics: vowel nasalization across a boundary. Ms., Johns Hopkins University.

Mullennix, John, David Pisoni, and Christopher Martin. 1989. Some effects of talker variability on spoken word recognition. *Journal of the Acoustic Society of America, 85*, 365-378.

Munson, Benjamin, and Nancy Pearl Solomon. 2004. The effect of phonological density on vowel articulation. *Journal of Speech, Language and Hearing Research, 47*, 1048-1058.

Nielsen, Kuniko. 2005. Generalization of phonetic imitation across place of articulation. In *Proceedings of the ISCA Workshop on Plasticity in Speech Perception*, University College London, London, UK.

Nielsen, Kuniko. 2006. Specificity and Generalizability of Spontaneous Phonetic Imitation. In *Proceedings of the 9th International Conference on Spoken Language Processing*, Pittsburgh, PA.

Nielsen, Kuniko. 2008. Word-level and Feature-level Effects in Phonetic Imitation. Doctoral dissertation, UCLA, Los Angeles, CA.

Nusbaum, Howard, David Pisoni, and Christopher Davis. 1984. Sizing up the Hoosier mental lexicon: measuring the familiarity of 20,000 words. *Research on Speech Perception: Progress Report No. 10*. Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Nygaard, Lynne and David Pisoni. 1998. Talker-specific perceptual learning in speech perception. *Perception & Psychophysics, 60*, 355-376.

Nygaard, Lynne. 2008. Perceptual integration of linguistic and nonlinguistic properties of speech. In David Pisoni and Robert Remez (eds.), *The Handbook of Speech Perception*, 390-413. Malden, MA: Blackwell.

Pardo, Jennifer. 2006. On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America, 119*, 2382–2393.

Port, Robert and Rosemarie Rotunno. 1979. Relation between voice-onset time and vowel duration. *Journal of the Acoustical Society of America, 66.3*, 654-662.

Scarborough, Rebecca. 2003. Lexical confusability and degree of coarticulation. In *Proceedings of the 29th Annual Meeting of the Berkeley Linguistics Society*.

Scarborough, Rebecca. 2004. Coarticulation and the Structure of the Lexicon. Doctoral dissertation, UCLA, Los Angeles, CA.

Shockley, Kevin, Sabadini, Laura, and Carol Fowler. 2004. Imitation in shadowing words. *Perception & Psychophysics, 66.3*, 422-429.

Smith, Rachel. 2007. The Effect of talker familiarity on word segmentation in noise. In *Proceedings of the 16th International Congress of Phonetic Sciences*, 1917-1920, Saarbrücken, Germany.

Swartz, Bradford. 1992. Gender differences in voice onset time. *Perceptual and Motor Skills*, 75, 983-992.

Whiteside, Sandra, Luisa Henry, Rachel Dobbin. 2004. Sex differences in voice onset time: a developmental study of phonetic context effects in British English. *Journal of the Acoustical Society of America*, 116.2, 1179-1183.

Wright, Richard. 1997. Lexical competition and reduction in speech: a preliminary report. *Research on Spoken Language Processing, 21*, 471-485. Bloomington, IA: Speech Research Laboratory, Psychology Department, Indiana University.

Wright, Richard. 2004. Factors of lexical competition in vowel articulation. In John Local, Richard Ogden, and Rosalind Temple (eds.), *Laboratory Phonology VI*, 26-50. Cambridge: Cambridge University Press.

Zue, Victor. 1980. *Acoustic Characteristics of Stop Consonants: A Controlled Study*. Bloomington, IN: Indiana Linguistics Club.

Proceedings of the 27th West Coast Conference on Formal Linguistics

edited by Natasha Abner and Jason Bishop

Cascadilla Proceedings Project Somerville, MA 2008

Copyright information

Proceedings of the 27th West Coast Conference on Formal Linguistics
© 2008 Cascadilla Proceedings Project, Somerville, MA. All rights reserved

ISBN 978-1-57473-428-7 library binding

A copyright notice for each paper is located at the bottom of the first page of the paper.
Reprints for course packs can be authorized by Cascadilla Proceedings Project.

Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, e-mail: sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Nielsen, Kuniko and Colin Wilson. 2008. A Hierarchical Bayesian Model of Multi-level Phonetic Imitation. In *Proceedings of the 27th West Coast Conference on Formal Linguistics*, ed. Natasha Abner and Jason Bishop, 335-343. Somerville, MA: Cascadilla Proceedings Project.

or:

Nielsen, Kuniko and Colin Wilson. 2008. A Hierarchical Bayesian Model of Multi-level Phonetic Imitation. In *Proceedings of the 27th West Coast Conference on Formal Linguistics*, ed. Natasha Abner and Jason Bishop, 335-343. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #1848.