A Bayesian Network Model of Multi-level Phonetic Imitation

Keywords: Phonetics, Learning, Bayesian

Introduction. Experiments in the imitation paradigm have established that speakers modulate the fine phonetic detail of their own productions in order to more closely approximate recently heard speech (e.g., Goldinger 1998, Nielsen 2007; see also Carlson et al. 2007, Dahan & Scarborough 2005, Maye et al. 2005). This imitation or 'phonetic convergence' effect has also been found in conversational interactions (Pardo 2006), suggesting that it is a normal and automatic part of phonetic processing. Previous formal analyses have focused primarily on the issue of whether storage of word-level exemplars of heard speech can account for the imitation effect. In this paper, we present an alternative analysis that exploits Bayesian computations in a model of the speech production system containing multiple levels of representation (word, segment, and gesture). The Bayesian analysis provides a quantitatively detailed and statistically interpretable account of the imitation effect for each of the participants in Nielsen's (2007) experiment on VOT. The essence of the analysis, which employs standard Bayesian learning techniques (e.g., Bishop 2006), is that listeners update their internal probabilistic phonetic models in response to perceived speech. The updated phonetic models probabilistically generate 'imitative' or 'convergent' productions.

Imitation experiment. Nielsen (2007) presented 27 participants with speech stimuli in which the voice onset time (VOT) of the word-initial consonant — always /p/ — was digitally extended by approximately 40ms. In comparison to *baseline* productions of a word list that were recorded prior to the exposure to VOT-extended speech, *test* recordings of the same list following exposure showed significantly longer VOTs. This imitation effect was found not only for items with initial /p/, but also for items beginning with /k/; furthermore, lexical influences on imitation of the kind reported by Goldinger (1998) were only weakly represented in this experiment (see also Carslon et al. 2007, Maye et al. 2005). The generalization from /p/ to /k/ and the relatively weak lexical effects are problematic for word-level exemplar models, and suggests that imitation emerges from phonetic processing and learning at multiple representational levels.

Bayesian network analysis. For the purposes of generating the VOT of a word-initial voiceless stop, we take the speech production system to consist of units at the word, segment, and gestural level (figure 1); additional units, responsible for prosodic or sociolinguistic conditioning, could be easily incorporated. Interpreting this system as a Bayesian network involves assigning a random variable to each node. The word and segment nodes follow discrete distributions; for convenience, we treat the gestural node as discrete as well (its value is [+spread glottis] in all tokens considered here). Conditioned on the higher nodes, VOT is assumed to follow a Gaussian distribution; to facilitate comparison of results across speakers, we converted measured VOTs to *z*-scores within participants — thus the role of higher nodes is to condition where a given token's VOT will fall within the standardized distribution of each speaker. The equation that expresses this relationship is: $\text{VOT} \sim N(w_{lex}*x_{lex} + w_{seg}*x_{seg} + w_{gest}*x_{gest}, 1)$, where each pair of w and x terms denotes the weight and mean value of one of the higher nodes. The term x_{seg} encapsulates two means, one for each relevant value of the segment node (/p/, /k/); x_{lex} includes one mean value for each of the 120 words in the experiment. These values, along with x_{gest}, were estimated from the baseline data for each individual speaker. Without loss of generality, all weights were initially set to 1.

Results. The baseline means and weights define a *prior* conditional distribution on VOTs. Exposure to expanded-VOT speech results in a *posterior* distribution that is derived by combining the prior with the likelihood of the exposure data in the way specified by Bayes' Theorem. To implement this learning process, we placed a Gaussian distribution on the weights with a common mean of **1** (i.e., a vector of all ones) and a precision parameter α for each speaker; informally, α controls the 'plasticity' or 'learning rate' of the system, and was the single parameter that we fit to each speaker's test productions. Standard results in the Bayesian literature imply that the posterior distribution on the weights and the predicted VOT distribution are also Gaussian (Bishop 2006: 152ff.). Figure 2 shows the predicted and observed values for each speaker's test productions, with a non-parametric regression line for each consonant superimposed on the scatterplot to aid interpretation. The positive correlations for nearly all speakers demonstrate the ability of simple, mathematically sound Bayesian models such as ours to account for the phonetic imitation effect. More detailed inspection shows that the multi-leveled model correctly predicts the degree of generalization from /p/ to /k/ and the weak but non-negligible effects of lexical identity.

A Bayesian Network Model of Multi-level Phonetic Imitation

Keywords: Phonetics, Learning, Bayesian

Figure 1: Bayesian network for VOT production

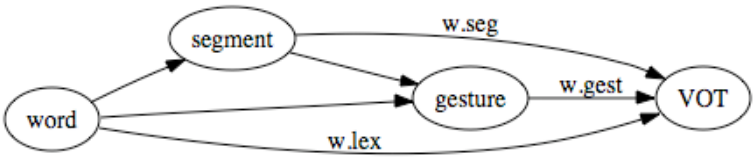

Figure 2: Predicted and observed standardized VOT values for each participant