True to Thyself: Assessing Whether Computational Models of Cognition Remain Faithful to Their Theoretical Principles

In Jae Myung (MYUNG.1@Osu.Edu)
August E. Brunsman IV (BRUNSMAN.3@Osu.Edu)
Mark A. Pitt (PITT.2@Osu.Edu)
Department of Psychology, Ohio State University
1885 Neil Avenue
Columbus, OH 43210 USA

Abstract

This study investigated the model selection problem in cognitive psychology: How should one decide between two computational models of cognition? The focus was on model "faithfulness, " which refers to the degree to which a model's behavior originates from the theoretical principles that it embodies. The guiding principle is that among a set of models that simulate human performance equally well, the model whose behavior is most stable or robust with variation in parameter values should be favored. This is because such a model is likely to have captured the underlying mental process in the least complex way while at the same time being faithful to the theoretical principles that guided the model's development. Sensitivity analysis is introduced as a tool for assessing model faithfulness. Its application is demonstrated in the context of two localist connectionist models of speech perception, TRACE and MERGE.

Introduction

One of the most challenging tasks for researchers interested in modeling human cognition is developing techniques for choosing among a set of computational models (e.g., Grossberg, 1987; Grainger & Jacobs, 1998). The goal is to choose the model that best captures the underlying cognitive process. It is standard practice to select the model that most accurately simulates or fits data generated by humans. Justification for using this procedure, termed descriptive adequacy, is that the best-fitting model most closely approximates the mental process being modeled. The adequacy of this model selection criterion is limited to cases in which the models do not capture the underlying process equally well. When they do, how should one decide among models? There are at least two important issues to consider.

The first issue is *model complexity*, which refers to the flexibility inherent in a model (i.e., how the parameters are combined mathematically) that enables it to fit diverse patterns of data. A model may describe data well, but may not do so in a parsimonious manner. It is well established (e.g., Linhart & Zucchini, 1986; Myung, in press) that model selection based solely on descriptive adequacy will result in the choice of an unnecessarily complex model that over-fits the data, and therefore fails to capture the true regularities of the underlying mental process. To avoid this mistake, both descriptive adequacy and model complexity must be taken into account in model selection (Myung & Pitt, 1997). These two criteria embody the principle of Occam's razor in model selection: The model that fits data sufficiently well in the least complex way should be preferred.

The second, equally critical, issue is determining the cause of a model's behavior. Is a model's success in mimicking human behavior due to the theoretical principles embodied in the model or due to other aspects of its computational instantiation? Put another way, is the computational instantiation faithful to its theoretical principles? These are not one in the same, as the latter can take on many forms (Uttal, 1990). Even if a model provides an excellent description of human data in the simplest manner possible, it is often difficult to determine what properties of the model are critical for explaining human performance and what aspects are not. Ideally, theoretical principles from which the model was developed must be clearly identifiable and their contribution to determining model behavior clearly demonstrated. In other words, the behavior of the model must originate from the theoretical ideas that motivated its creation, not from the computational choices made in its instantiation. Failure to make this distinction runs the risk of erroneously attributing a model's behavior to its underlying theoretical principles: computational complexity is mistaken for theoretical accuracy.

In this paper, we undertake an investigation of model faithfulness. The behaviors of two localist models of phonemic perception, TRACE (McClelland & Elman, 1986; McClelland, 1991) and MERGE (Norris et al, 1998) were compared to determine which architectural properties are most responsible for their behavior. Sensitivity analysis, a measure of how sensitive the behavior of a model is to variation in the values of its parameters, is introduced as a tool to assess model faithfulness. An additional attractive

property of sensitivity analysis is that the results reveal the relative complexities of the models. A highly sensitive model is complex. A small change in the value of a parameter can change the model's behavior drastically. This property makes the model very powerful and adaptable, being able to fit a wide range of data patterns, perhaps many more than are necessary to model the mental process of interest. On the other hand, a model whose behavior changes minimally with variation in parameter values is far less flexible, but behaviorally more stable (i.e., robust). If such a model happens to simulate human data well, then it is an indication that the model may have captured the regularities of interest in the data and little else. Following Occam's razor, such a model should be favored. Thus, model selection using sensitivity analysis favors the model that is least sensitive to parameter variation, and as a consequence captures the data the best under the widest range of parameter variation.

Connectionist Models of Speech Perception

Model development in speech perception is divided on the issue of how prior information from different sources is integrated during recognition (see Frauenfelder & Tyler, 1987). A wide range of experimental results has demonstrated that a listener's knowledge about a word can influence how the phonemes (i.e., speech sounds) of that word are perceived. The theoretical debate in the literature has focused on determining how these two forms of information (lexical and phonemic) are combined during perception. In most computation models of word recognition, there exist at least two levels of processing, a phonemic level and a lexical (i.e., word) level. Information flow from the phoneme to the lexical level is common among models. The theoretical distinction of primary importance is how lexical information is integrated with phonemic information. In Figure 1, TRACE and MERGE illustrate the two positions architecturally. In TRACE (McClelland & Elman, 1986), activation of phonemes is influenced by bottom-up sensory input from the speech signal itself *and* from top-down connections to the lexical level. In MERGE (Norris et al, 1998), there are no top-down connections from the lexical level that *directly* affect phoneme activation. Rather, phonemic processing is split in two, with an activation/input stage and a phonemic decision stage. Lexical processes affect only phonemic decision making; they cannot directly influence phoneme activation. In MERGE, phonemic and lexical influences on phonemic decision making are independent of each other, being integrated only at the decision stage.

Norris et al (in press) showed that the models are fairly comparable in their ability to simulate two sets of human data. But what properties of the models are most responsible for their similar behavior? MERGE was proposed as a non-interactive alternative to TRACE, with no top-down feedback directly to the phonemic input stage. However, the models

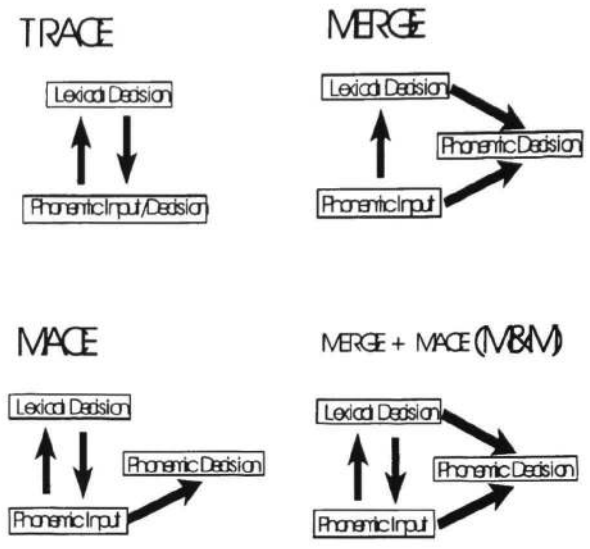

Figure 1. Network architectures of the four models tested.

also differ in the number of phonemic stages, with TRACE possessing one and MERGE possessing two. Which of these properties is most responsible for MERGE's behavior, direction of information flow or an additional stage? Additionally, are the two models equally complex? That is, we know their architectures are sufficient to capture the test data, but are they also necessary? Or are there redundancies in their design that make them overly complex?

To address these questions, sensitivity analyses were carried out on TRACE and MERGE and two other models (shown in the bottom of Figure 1), which were included to understand better the implications of model design on model performance. MACE is a hybrid of TRACE and MERGE that was intended to assess the consequences of independently integrating lexical and phonemic input at a separate decision stage. Like TRACE, lexical information flows back to the phonemic activation level, but like MERGE, phoneme decisions are made separately. If MERGE derives its descriptive power solely from its non-interactive architecture, then MACE's performance should be significantly inferior to that of MERGE. Similar performance would suggest that splitting phoneme processing into two stages is more important than whether lexical information flows directly to the phoneme activation level or instead to the phoneme decision level. MERGE+MACE (M&M) is a combination of MERGE and MACE. Lexical connections are redundant, feeding to both the phoneme activation level and the decision making level. It is included for completeness and to serve as a check on the predictability of models with different configurations of information flow.

Table 1. Experimental Conditions and Human Data.

Experimental Condition		Human Data					
Condition	Example	Phonemic				Lexical	
		/b/	/g/	/z/	/v/	"job"	"jog"
W1W1 (Word 1 + Word 1)	JOb + joB (JOB)	*				*	
W2W1 (Word 2 + Word 1)	JOg + joB (JOB)	*				*	
N2W1 (Nonword 2 + Word 1)	JOv + joB (JOB)	*				*	
N1N1 (Nonword 1 + Nonword 1)	JOz + joZ (JOZ)			*			
W2N1 (Word 2 + Nonword 1)	JOg + joZ (JOZ)			*			
N2N1 (Nonword 2 + Nonword 1)	JOv + joZ (JOZ)			*			

Note: For each condition, the star '*' indicates the phoneme or word that is recognized by listeners.

Method

Overview

Norris et al (in press) evaluated TRACE and MERGE on their ability to simulate data showing listeners' sensitivity to mismatching phonemic information at the end of an utterance (Whalen, 1984; Marslen-Wilson & Warren, 1994). This same data set was used in the evaluation of the four models in Figure 1. First, the ability of the models to simulate the human data was assessed to replicate Norris et al and demonstrate that all models were comparable in descriptive adequacy. Second, a sensitivity analysis was performed on the models by systematically varying the parameter values around the optimal parameter settings that provided the best fit to the data in the first analysis. As mentioned above, the sensitivity analysis assessed the robustness of a model's behavior in the face of parameter variation. It enabled us to ascertain the degree to which performance arises from theoretical principles that the model purports to implement. The more the behavior of a model changes over the range of parameter values, the less likely the model derives its power from its theoretical design principles, in this case how lexical and phonemic information are integrated, than from idiosyncratic choices of parameter values.

Data That Were Modeled

Following Norris et al (in press), the four models were compared in their ability to simulate data from Marslen-Wilson and Warren (1994; McQueen, Norris & Cutter, in press), in which listeners were shown to be sensitive to conflicting phonemic input in both phonemic decision making and lexical decision making. When listening to speech, listeners exhibit considerable sensitivity to deviations from the natural production of an utterance. For example, if the portion of the phoneme (i.e., letter sound) /g/ in the word "jog" is spliced off and replaced with a token of the phone /b/ from the word "job," listeners are slower to identify the final phoneme, /b/, in the newly created cross-spliced word "job" than in the original, unspliced token of "job." This is because the acoustic information signaling the identity of the final phoneme cannot be fully removed, as it blends into the immediately preceding vowel, creating conflicting information about the identity of the final phone. The acoustic information at the end of the vowel specifies /g/ whereas the subsequent information specifies /b/.

By varying the original source of the two parts of a cross-spliced token (i.e., whether they came from words or nonwords), lexical influences on phoneme and word processing can be explored. The six conditions shown in Table 1 were used. The alpha-numeric condition names (column 1) refer to the composition of the cross-spliced stimulus. For example, W1W1 refers to the stimulus described above in which the cross-spliced stimulus was created from two words. In the examples, the capital letters refer to the portions of the two utterances that formed part of the cross-spliced utterance, with the resulting cross-spliced stimulus in parentheses. Both identification of the final phoneme and recognition of the cross-spliced word were simulated. For each condition, the asterisk '*' indicates the phoneme or word that was recognized by listeners.

Phonemic Decision Making Data

W1W1 and N1N1 were two control conditions, included only to demonstrate that when no conflicting phonemic cues are present in the stimulus, recognition of the final phoneme is not impeded. In the four remaining conditions, there is conflicting phonemic information in the cross-spliced tokens because stimuli with different final phonemes were cross-spliced. The lexical status of the two "source" utterances influences identification of the final phoneme. When the initial item is a word (e.g., "jog"), as in conditions W2W1 and W2N1, recognition of the final phoneme, /b/ or /z/, are comparatively slower, presumably because the lexical entry for "jog" affects phonemic processing in some manner. Further, the amount of the slowdown is less in W2W1 than in W2N1, which is thought to be due to lexical competition between "job" and "jog" diminishing lexical influences in processing the final phoneme. If the initial portion of the cross-spliced stimulus comes from a nonword (e.g., "jov'), as in conditions N2W1 and N2N1, there is no slowdown in recognition of /b/ or /z/. Although there is conflicting phonemic information in the cross-spliced stimulus, the use of a nonword effectively shuts down lexical influences.

Table 2. Description of Model Parameters

Parameter	TRACE	MERGE	MACE	M&M
PE (phoneme excitation)	✔	✔	✔	✔
PWE (phoneme to word excitation)	✔	✔	✔	✔
PWI (phone to word inhibition)		✔	✔	✔
PTE (phoneme to target excitation)		✔	✔	✔
PD (phoneme decay)	✔	✔	✔	✔
WTE (word to target excitation)		✔		✔
WPE (word to phoneme excitation)	✔		✔	✔
WWI (word to word inhibition)	✔	✔	✔	✔
WD (word decay)	✔	✔	✔	✔
TTI (target to target inhibition)		✔	✔	✔
TD (target decay)		✔	✔	✔
TM (target momentum)		✔	✔	✔
CPS (word/target cycles per input slice)	✔	✔	✔	✔
PPI (phoneme to phoneme inhibition)	✔			

Note: For each parameter, the check '✔' indicates the models that adopt the parameter.

The bottom-up information specifying /v/ is too weak to affect phonemic decision making.

Lexical Decision Making Data

Lexical decision making with the cross-spliced stimuli was straightforward. Between the two possible word responses (e.g., "job" and "jog"), if the final item in the cross-splice is a word (e.g., "job") as in conditions W1W1, W2W1 and N2W1, then only the word "job" (i.e., final item) should be recognized, regardless of whether the initial portion originally came from a word or a nonword. On the other hand, if the second portion came from a nonword (e.g., /z/ in "joz") as in conditions N1N1, W2N1, and N2N1, then word recognition depends on whether the initial item in the cross-splice is a word or nonword. If it is a word (e.g., "jog") as in W2N1, then the word "jog" should be activated, but not substantially to be recognized because of the following mismatching information (/z/). If it is a nonword (e.g., "joz" or "jov") as in N1N1 and N2N1, then both "jog" and "job" will be activated too weakly to be recognized.

Model Implementation and Simulation Procedure

The four models in Figure 1 were constructed by modifying the architecture of MERGE, which is a localist network consisting of six input nodes corresponding to the phonemes /ʤ/ "j", /o/, /b/, /g/, /v/ and /z/, four lexical decision nodes representing two words ("job", "jog") and two nonwords ("jov", "joz"), and finally, four phonemic decision nodes representing the target phonemes /b/, /g/, /v/ and /z/. In MERGE, each lexical decision node receives inputs from the phonemic input nodes through excitatory connections and also receives activations from other lexical decision nodes through lateral inhibitory connections. Similarly, the phonemic decision nodes are linked to the phonemic input nodes as well as to the lexical decision nodes through excitatory connections. Lateral inhibitory connections are assumed among phonemic decision nodes whereas no such connections are assumed among phonemic input nodes. TRACE, MACE, and M&M, were created either by pruning existing connections and/or adding new connections to MERGE. The TRACE model had 8 parameters, MERGE and MACE models had 12, M&M model had 13 parameters. The parameters used in the models are cross-tabulated in Table 2 to provide one view of their similarities and differences.

In simulating the human data, each model was presented with the same input, six numerically represented cross-spliced tokens that were all three phonemes long and were either words or nonwords. All six of the tokens, one for each condition, were identical to the ones used by Norris, McQueen, and Cutler (1998). Each token was represented by six Mx1 vectors (one vector for each phoneme) where M is the number of time slices or iterations. The first vector gave activation to the /ʤ/ node in the phoneme input layer, the next to /o/, and so on for /b/, /g/, /v/, and /d/. Each vector began at zero except for /ʤ/, which was .25 in the first time slice, .5 in the second, and then 1 in the third (maintained for the rest of the iterations). In the fourth time slice, /o/ went to .25, .5 in the fifth, and 1 in the sixth (maintained for the rest of the trail). In a similar fashion, the final phoneme was constructed, depending upon the condition. For a given input stimulus, the activation profiles of the phonemic decision nodes and the lexical decision nodes were obtained and then compared to the predictions from human data to evaluate the model's performance.

Each of the four models was qualitatively fit to the human data, and a set of optimal parameter values was obtained by a hand-done parameter search, relying on initial estimates reported in Norris, McQueen, & Cutler (1998). A model's

behavior was judged to be either "human-like" or "not human-like"in each condition by determining whether the model's output matched predictions in the corresponding condition. Twelve judgements were made for each model, six phoneme decisions and six lexical decisions (Table 1). In the sensitivity analysis, the parameter values of each model were systematically varied ±75% from the optimum value. At each value of the parameter, the model's behavior (phonemic decision making and lexical decision making) was re-assessed in the twelve judgements.

Results and Discussion

Simulating Human Data

All four models produced "human-like" results in every condition in both phonemic decision making and lexical decision making, including the all-important slowdown in phonemic processing in conditions W2W1 and W2N1. Thus in terms of descriptive adequacy, all models were functionally equivalent in their ability to simulate this set of data. This finding suggests that the two ways in which lexical and phonemic information are integrated does not matter: Direct top-down feedback (TRACE, MACE) simulates human performance just as well as integrating the two sources of information independently (MERGE) or a combination of the two methods (M&M).

Given the similar behavior of the four models, how should we choose among them? Overly complex models should be avoided. Recall that MERGE, MACE and M&M assume that phoneme activation is separate from phoneme decision making. Relative to TRACE, which makes no such distinction, these models require extra parameters (4 for MERGE and MACE, 5 for M&M). The finding of virtually no difference in descriptive adequacy between any of the models suggests that splitting phonemic processing across levels is a redundant property of these models, one unnecessary to simulate human behavior. Instead, splitting phonemic processing in two may introduce unnecessary complexity that only reduces generalizability of the models, making them less stable amidst parameter variation. The sensitivity analysis explores this possibility.

Sensitivity Analysis

Figure 2 shows the proportion of non-human-like data patterns (i.e., errors) generated by each model when the model's parameters were systematically varied around the optimum values. The proportions were averaged over all parameters and are shown separately for each of the 12 testing conditions. TRACE was the least error prone, with a 2.1% error rate, whereas the other three models made considerably more errors (5 - 7 %). This result is clear confirmation that splitting phonemic processing into two stages (MERGE, MACE, and M&M) does more than reflect the regularities of human speech processing. It introduces

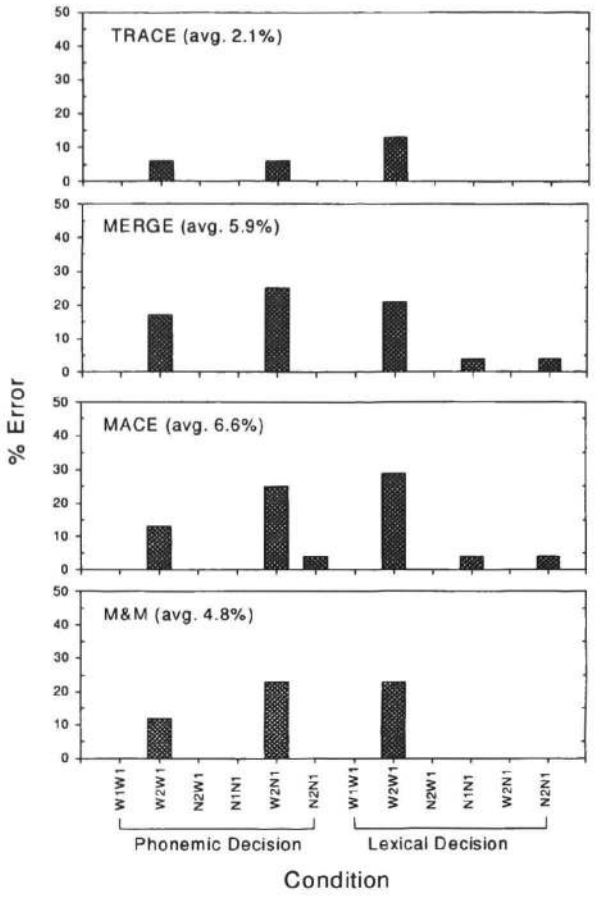

Figure 2. Results of sensitivity analysis..

unnecessary flexibility not needed to capture the phenomena of interest.

Examination of Figure 2 reveals that errors occurred most frequently in three conditions: W2W1, W2N1 in phonemic decision making and W2W1 in lexical decision making. In the first two, phonemic and lexical information must be integrated to simulate accurately human data. In the third, the effects of lexical inhibition must be simulated. The three models with independent lexical and phonemic influences on phonemic decision making (MERGE, MACE, M&M) were very sensitive to parameter variation in these conditions, producing many patterns that were not human-like. TRACE, on the other hand, was able to exhibit human-like performance under a wider range of parameter values.

Note also in Figure 2 the similar performance profiles of MERGE and MACE across the 12 conditions. Recall that MACE, like TRACE, contains top-down flow of lexical information directly to the phonemic input level, but like MERGE, phonemic processing is split into two stages. The fact that this hybrid model behaves so similarly to MERGE suggests that MERGE's behavior is determined more by the separation of phonemic processing into two stages than by lexical information affecting phonemic decision making rather than phonemic activation. In other words, what

differentiates MERGE from TRACE is not so much how information flows between processing stages, but the number of processing stages. This result suggests that the current implementation of MERGE is only partially faithful to the theoretical principle that motivated its development.

The sensitivity analysis suggests that the extra parameters that MERGE, MACE, and M&M require as a result of separating phonemic processing into two stages, and thus making the models non-interactive, increases the complexity of the models. This design characteristic has the detrimental side effect of decreasing model robustness. TRACE explains human data sufficiently well in the least complex manner.

Summary and Conclusion

The purpose of this preliminary investigation was to explore the model selection problem in cognitive psychology: How should one decide between two computational models of cognition? The particular focus of the study has been on assessing model faithfulness, which refers to the degree to which a model's behavior originates from the theoretical principles that it embodies. The idea is that among a set of models that simulate human performance equally well, the model whose behavior is most stable with variation of parameter values should be favored. This is because such a model is likely to be most faithful to the theoretical principles that guided the model's development; it is also likely to have captured the underlying mental process in the least complex way. Sensitivity analysis was introduced as a tool for assessing model faithfulness. An application of the method was demonstrated for comparing the behaviors of four connectionist models of speech perception, TRACE, MERGE, MACE and M&M.

All four models were functionally indistinguishable in their ability to simulate human data. Sensitivity analysis, however, revealed that TRACE was the most stable model, suggesting that it best reflects the underlying regularities of human behavior and therefore should be preferred. An important implication of these results for modeling speech perception is that the separation of phonemic decision making from phonemic activation, as assumed in MERGE, MACE, and M&M, may be an overly complex architectural design that is not necessary to capture the phenomenon of interest (i.e., lexical and phonemic interaction).

Acknowledgments

The authors wish to thank Dennis Norris for kindly answering many questions we had about implementation of MERGE. This research was supported by NIMH Grant MH57472 to I.J.M. and M.A.P., and by the Ohio State University Colleges of Arts and Sciences Honors Award to A.E.B.

References

Frauenfelder, U.H. & Tyler, L.K. (1987) The process of spoken word recognition: An Introduction. *Cognition* 25:1-20.

Grainger, J., & Jacobs, A. M. (1998). On localist connectionism and psychological science. In J. Grainger & A. M. Jacobs (eds.), *Localist Connectionist Approaches to Human Cognition*. Lawrence Erlbaum Associates.

Grossberg, S. (1987). Competitive learning: From interactive activation and adaptive resonance. *Cognitive Science*, 11, 23-63.

Linhart, H., & Zucchini, W. (1986). *Model Selection*. New York: Wiley.

Marslen-Wilson, W. & Warren, P. (1994) Levels of perceptual representation and process in lexical access: words, phonemes, and features. *Psychological Review* 101:653-675.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18, 1-86.

McClelland, J. L. (1991). Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology*, 23, 1-44.

McQueen, J. M., Norris, D. & Cutler, A. (in press) Lexical influence in phonetic decision making: Evidence from subcategorical mismatches. *Journal of Experimental Psychology: Human Perception & Performance.*

Myung, I. J. (in press). The importance of complexity in model selection. *Journal of Mathematical Psychology.*

Myung, I. J., & Pitt, M. A. (1997). Applying the Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, 4, 79-95.

Norris, D., McQueen, J.M. & Cutler, A. (in press) Merging phonetic and lexical information in phonetic decision-making. *Behavioral and Brain Sciences.*

Uttal, W. R. (1990). On some two-way barriers between models and mechanisms. *Perception & Psychophysics*, 48, 188-203.

Whalen, D. H. (1984). Subcategorical phonetic mismatches slow phonetic judgments. *Perception & Psychophysics*, 35, 49-64.