A Dynamic Neural Field Model of Speech Cue Compensation

Gavin W. Jenkins (gjenkins@sfu.ca)
Paul Tupper (pft3@sfu.ca)
Department of Mathematics, 8888 University Drive
Burnaby, B.C., V5A 1S6 Canada

Abstract

Categorical speech content can often be perceived directly from continuous auditory cues in the speech stream, but human-level performance on speech recognition tasks requires compensation for contextual variables like speaker identity. Regression modeling by McMurray and Jongman (2011) has suggested that for many fricative phonemes, a compensation scheme can substantially increase categorization accuracy beyond even the information from 24 un-compensated raw speech cues. Here, we simulate the same dataset instead using a neurally rather than abstractly implemented model: a hybrid dynamic neural field model and connectionist network. Our model achieved slightly lower accuracy than McMurray and Jongman's but similar accuracy patterns across most fricatives. Results also compared similarly to more recent models that were also less neurally instantiated but somewhat closer fitting to humans in accuracy. An even less abstracted model is an immediate future goal, as is expanding the present model to additional sensory modalities and constancy/compensation effects.

Keywords: Speech recognition, concepts and categories, neural networks, dynamic systems modeling, psychology, linguistics, cognitive science

Constancy and Compensation in Perception

In most contexts, our senses provide more information than we require for a decision. This can make recognition tasks difficult when the undesired, noisy information is not just alongside but integrally mixed with desired information. As examples, the overall lighting of a scene as well as the reflectance or color of an object both affect the object's perceived lightness and hue; the actual shape of an object and viewing angle both affect perceived shape; and a speaker's gender and the content of his or her speech both affect sound pitch. Humans are adept at discounting noise and ambiguities, achieving location, shape, or speech cue constancy (Schneegans & Schöner, 2012; Rock, 1983; Bendor & Wang, 2005). Here, we present a neurally plausible, computational model potentially suitable for any type of constancy that relies on discounting dimensional feature information such as hue, shape cues, or speech cues.

Specifically, we test the model by identifying fricative consonants ('fricatives') from whole spoken syllables. We assume a speech cue to phoneme model, in contrast to a more purely acoustic approach (Graves, Mohamed, & Hinton, 2013; Pisoni, 1997), but our model could adapt to a raw acoustic approach with very similar architecture.

In the speech cue approach, few, if any, speech cues for fricatives are considered "invariant." That is, individual cues like vowel duration do not statically, cleanly, and conveniently correlate with phoneme categories (though there is some debate, see Stevens & Keyser, 2010). Rather, most or all cue information shifts contingently based on the contextual vowel sounds, speed of speech, or speaker.

Recent empirical and modeling evidence suggests that cue invariance can be overcome by considering very large numbers (dozens) of speech cues and by actively identifying contexts then normalizing incoming speech compared to other contexts. Jongman, Wayland, and Wong (2000) gathered human listener identification data for eight fricatives (f, v, θ, ð, s, z, ʃ, ʒ) or henceforth (f, v, th, dh, s, z, sh, zh) respectively, ranging from labiodental to post-alveolar place of articulation and including both voiced and voiceless fricatives at each place. Recordings of fricatives spanned over 20 speakers and 6 vowel contexts. McMurray and Jongman (2011) then tested this corpus of data using an abstract logistic regression model. They tested the model under several learning conditions, including a small collection of 10 fricative-only cues, a large set of 24 cues that also added vowel cues, and the same 24 cues, but with expectation-based context compensation for vowel and speaker, using the formal regression compensation model C-Cure (McMurray, Cole, & Munson, 2011). They determined that higher numbers of cues contributed to more accurate identification per fricative, as did expectation-based context compensation. Context compensation also allowed the model to fit human behavior more closely.

Neural Implementation of Cue Compensation

McMurray and Jongman's (2011) results are a compelling demonstration of the importance of a large number of cues and of cue compensation for fricative identification. However, the model is abstract and mathematical in both phoneme categorization and cue compensation. Such models are important, but a model in a neural framework offers a chance to discover and understand processes driving behavior that may derive from neural-level interactions not considered at an abstract level. Formal neural models also generate testable and informative neural level tests and predictions. Apfelbaum and McMurray (2015) presented a neural two layer PDP neural network for phoneme categorization, which performed impressively and comparably to McMurray and Jongman's (2011) results, but cue representation, context identification and context compensation were all still abstracted mathematically.

Here, we present a neural model to capture the same behavioral data as McMurray and Jongman (2011) and to further expand and supplement our understanding of speech cue compensation from a neural perspective. We use a

dynamic neural field (DNF) model for attention, memory, and storage of known speaker speech profiles instead of direct coding of these steps into the model, a DNF cue compensation mechanism instead of C-Cure, and a single layer neural network to ultimately decide phonemes.

The DNF architecture is described in detail below, but in general, DNF models involve fields of neural units whose receptive fields are systematically organized by dimensional information like space, size, color hue, or pitch. The DNF approach does not assume that all cognition is organized this way, only a subset of representations and processing that involve dimensional data. This includes attention to certain dimension values (attending to particular colors or points in space, for example), memory traces of those values, or in the current model, shifting cue values along a dimension to compensate for speaker identity. Beyond these processes, connectionist networks often take control. Hybrid models can include both in simulations, such as McMurray, Horst, Toscano, and Samuelson (2009) or the current model.

DNF models have been used for simulating processes ranging from word learning (McMurray, et al., 2009; Samuelson, Spencer, & Jenkins, 2013), to motor planning (Erlhagen & Schöner, 2002), to object recognition (Faubel & Schöner, 2008) and more, and the current model uses many of the same neural fields as do the above models in the same layouts. Mechanisms for different cognitive processes provide testable predictions for one another and can potentially be considered together as a coherent whole and unified model. This is not an advantage exclusive to neural models, but it is natural to them, since a shared, fundamental language is encouraged by common neural level simulation.

Our model of fricative perception utilizes several already-established DNF and connectionist mechanisms. The core neural field dynamics are common to all DNF models; the perceptual and memory portions of the model are common to DNF models that involve categorization; the phonetic cue compensation mechanism is inspired by a spatial transformation mechanism used in a DNF model of head and eye gaze spatial adjustment (Scheegans & Schöner, 2012); and the categorization step is performed by a sigmoidal connectionist network rather than logistic regression, similar to Apfelbaum and McMurray (2015).

Dynamic Neural Field Model Architecture

The Dynamic Neural Field (DNF) model consists of many 1- and 2-dimensional fields, shown as white rectangles in Figure 1. Units in fields are organized by one or two metric dimensions like cue values (such as voice onset time) or amount of adjustment needed to a cue. Each unit in a field has a receptive field maximally sensitive to one input value along its metric dimension(s) and less so to nearby values.

Units within a field interact with one another, sending close, strong local excitation and weaker, more diffuse lateral inhibition to neighbors. Both these interactions and input receptive fields create Gaussian "peaks" and "ridges" of activation when a field is given even a single value of input. These interactions keep peaks stable and localized yet robust against noise. Fields can be parameterized to have peaks collapse once input is removed (such as for attention) or to sustain themselves afterward (for memory).

Fields interact along shared dimensions. A field organized by pitch might send activation from above-threshold peaks to contribute to corresponding peaks in other pitch fields. A 1-dimensional field projecting to a 2-dimensional one projects a "ridge" of activation across all units in the larger field with the corresponding receptive fields and vice versa.

Fields also receive spatially correlated noise "input." This is insufficient to form peaks of activation alone, but is able to meaningfully influence other, stronger activity.

Figure 1 shows the full layout of *just one speech cue* in the model. That is, all of Figure 1 is repeated in the model for every speech cue. Initial input arrives at the input cue field (black star icon, blue region). In the example shown, the listener is hearing two different values of a cue. This could be due to hearing two speakers simultaneously, for example. This input projects to the adjacent attention field. Attention is a competitive field, where above-threshold units project global inhibition to the rest of the field, leading to a winner-takes-all activation pattern.

Both attention and input fields then project activation to a working memory (WM) field. This is a field with self-sustaining peaks, holding information for a time even after it has died away in temporary perception or attention fields. WM connects to a number of other fields of units representing long term memory (LTM) of speech cues. (right side of Figure 1). LTM fields are not dynamic; they are feed-forward and activated in a 1:1 correspondence with working memory. Long term memory information is held in Hebbian connection weights between WM and LTM fields. Whenever peaks are active in WM fields, LTM units represent the mathematical product of recent memory activation and LTM patterns. Thus, the total activation in LTM fields is effectively a similarity rating between recent speech cues and long term remembered patterns. If each LTM field and its Hebbian weights hold information of the history of cue values of an individual speaker, this similarity signal allows a listener to identify a known speaker's voice by competitively comparing the summed activation of LTM fields (top right of Figure 1).

Using speaker identity from above, the model activates a corresponding memory of the correct adjustment for that speaker's irregularities (green region, Figure 1). The 2-dimensional transform field (green region) accepts input from this adjustment value (top) and from the raw attended value (right) and adds them into a normalized cue value (lower left). Addition is performed by the overlap in activation between raw and adjustment ridges creating an intersecting diagonal ridge that projects to the summed value in the lower left diagonal neural field (red region of Figure 1). The same mechanism is suggested by Schneegans and Schöner (2012) for adding the angle of head rotation and angle of eye gaze to determine angles between body and objects in the visual environment. When an adjustment is

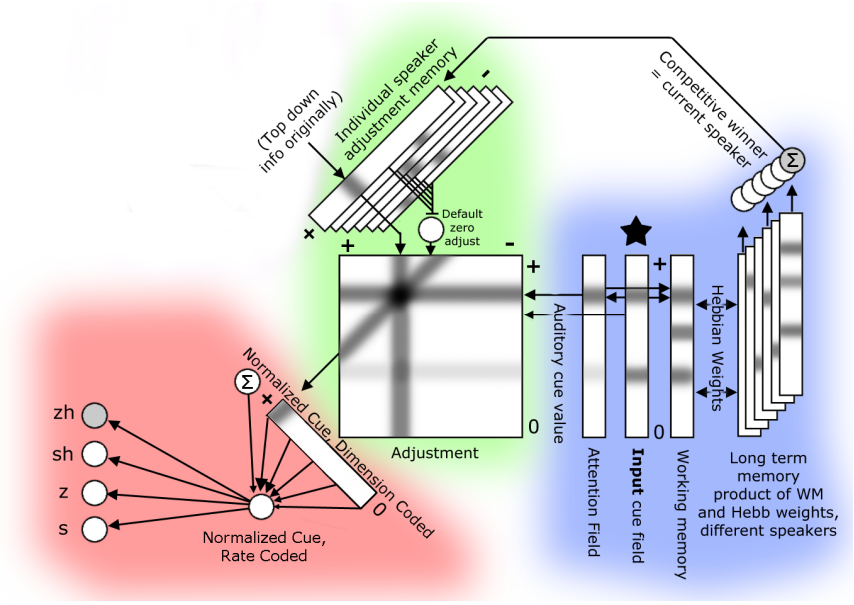

Figure 1. Architecture of a single cue in the DNF model. The full model consists of either 10 or 24 copies of this entire figure, except for just one set of output nodes shown in the lower left corner. See text for detailed description.

not known, a default adjustment of zero is used (suppressed otherwise).

The now-normalized cue information is transformed from dimensionally coded to rate coded format (red region of Figure 1). A gradient of connection weights projects stronger activation to the rate unit for peaks on one side of the cue dimension than the other. The sum of the dimensionally coded field also projects to the rate unit, allowing it to distinguish between no peak and a peak at the weak end of the scale. An almost identical neural circuit for place-to-rate code conversions is suggested by Groh (2001).

Finally, the normalized, rate-coded cue information feeds across a single layer network to determine the model's best guess at a phoneme. This final portion of the model equates to the neural portion of Apfelbaum and McMurray's 2015 PDP model of this data. Although only one cue node is shown, recall that all of the Figure 1 architecture is repeated *per cue*, creating a full [cues] x [phonemes] single layer network with one set of weights across speakers.

Experimental Design

We tested the DNF model using the fricatives phoneme dataset from Jongman, Wayland, and Wong (2000), including 8 fricatives spoken by 20 speakers each, in 6 vowel contexts (fricative + vowel syllables). McMurray and Jongman (2011) analyzed this data in several ways, but we focus here on neurally replicating three analyses in particular: phoneme categorization with 10 fricative-only cues and no compensation; with 24 cues to both fricative and accompanying vowel with no compensation, and with 24 cues and compensation. Together, these test the importance of number of cues and vowel information (10 vs 24) and compensation (24 without vs 24 with) for accurate (and human behavior-fitting) phoneme categorization.

Differences in DNF model architecture compared to McMurray and Jongman's (2011) model necessitated some lesser complexity in practical simulations. The DNF model compensated only for speaker context, not vowel context, due to the ability to only perform one compensation in a transform field at a time. Two or more compensations could simply and plausibly be performed in two or more fields in parallel, but would not better prove the concept initially and would take much longer to simulate. The neural transform field is also only capable of linear shifting adjustments along a dimension, which is less theoretically sophisticated than C-Cure compensation.

The neural network (red) portion of the DNF model was tuned first, without dynamic neural field input. Dynamic fields involve highly parallel processing, and are thus unrealistically slow to simulate on computers. Dynamic fields were therefore switched on during testing only. Training of the network used two out of every three syllable tokens from the same dataset categorized by humans, pre-coded for speech cues. In each epoch of training, the network received all training tokens once, blocked by speaker. The network received 2,000 epochs of training in each condition, with a learning rate of 0.3 and sigmoidal activation function. Cues were (mathematically during training) compensated during training for speaker prior to the neural network, for conditions including compensation.

The DNF long term memory fields were also pre-loaded with memory traces matching each speaker's cue profile (representing our memories of specific people's voices) and each speaker's adjustment value. Adjustments were chosen such that a linear shift would cause each speaker's mean value in a cue across recordings to equal the population mean value in that cue among the whole set of speakers. The model could capably establish this information itself with DNF during training, but this was impractically slow.

At test, the entire model was connected and used as a whole to categorize one epoch of the reserved generalization tokens, using the previously trained neural network.

We recorded accuracy across test trials using two different choice rules, as did McMurray and Jongman (2011). A discrete-choice rule always chose the phoneme with the highest activation in the output of the decision network. A probabilistic rule treated relative activation of each output node as relative probability of choosing that phoneme. The results figure (Figure 2) depicts colored regions bounded by these two different measures.

Results

Results of simulations are shown in Figure 2. Compared to McMurray and Jongman's (2011) mathematical model, the neural implementation of phoneme perception performed somewhat less absolutely accurately and somewhat less well fitted to human performance across conditions. However, the simulations were overall comparable while including many neurally-implemented mechanisms previously only abstractly implemented.

A large number of cues including vowel cues (24 cues, middle panel) provided moderate benefit compared to a medium number of fricative-only cues (10 cues, top panel) in terms of raw accuracy (+8% McMurray and Jongman [M&J], +11% DNF). Fit to human data worsened for M&J (RMSE from 0.077 to 0.099). Fit remained steady for the DNF model (RMSE of 0.156 in both cases), but results broadly shifted in line with human results in both models.

The addition of a speaker compensation system boosted raw accuracy in both models (+8% M&J, +3% in the DNF model) and fit human data more closely for M&J (RMSE from 0.099 to 0.061) and more closely for the DNF model (RMSE from 0.1563 to 0.1231).

The bottom panel in Figure 2 includes two additional lines, representing results from Apfelbaum and McMurray (2015) models. The grey line shows the performance of their PDP neural decision network (but otherwise not neurally implemented) model. The PDP model performed well quantitatively but with a flat performance across fricatives. The green line shows the performance of an exemplar model that stored every individual syllable token from the training set in memory for use in categorizing test items. The exemplar model performed very well, but had no neural implementation, and storage of every individual syllable in memory is likely unrealistic.

Qualitatively, the pattern of accuracy in the DNF model fits human data in shape about as well as the other models. Non-sibilant fricatives 'f' and 'v' were routinely low accuracy compared to humans for the DNF model, but the 'th, dh, s' dip shape is more accurately captured across conditions by the DNF model.

In an attempt to explore possible causes of our model's consistently low accuracy for 'f' and 'v', we tested the model with portions of the data including only 2 phonemes and 1 cue at a time. Two examples of 2-phoneme, 1-cue results are shown in Figure 3. The top row is data before speaker compensation, and the bottom row after. Horizontal position is value along the listed cue dimension (measured from the normalized DNF cue field) and each thin rows of dots is a different speaker. For some cues like spectral kurtosis (left), linear speaker compensation was able to actually *lower* accuracy in this analysis. One speaker may have had higher values for 'v' than 'zh' while another had not just a shifted but an *opposite* relationship between the same phonemes. Thus, when speakers' adjustments were chosen to match global means, some values for each phoneme shifted one direction, others shifted another, and confusion between the phonemes actually increased!

For other cues, like fricative duration (right side of Figure 3), there was a consistent relationship across speakers for one phoneme's values versus another, and compensation helped across all speakers.

Running the DNF model without the worst cues did not, however, increase model fits. The best results (by a small margin) were found when the worst nine cues were left uncompensated, but the overall improvement was not significantly better fitting than when compensating all cues, suggesting the network in the model is capable of discounting unproductive cues sufficiently. Low DNF accuracy for 'f' and 'v' fricatives remains unexplained.

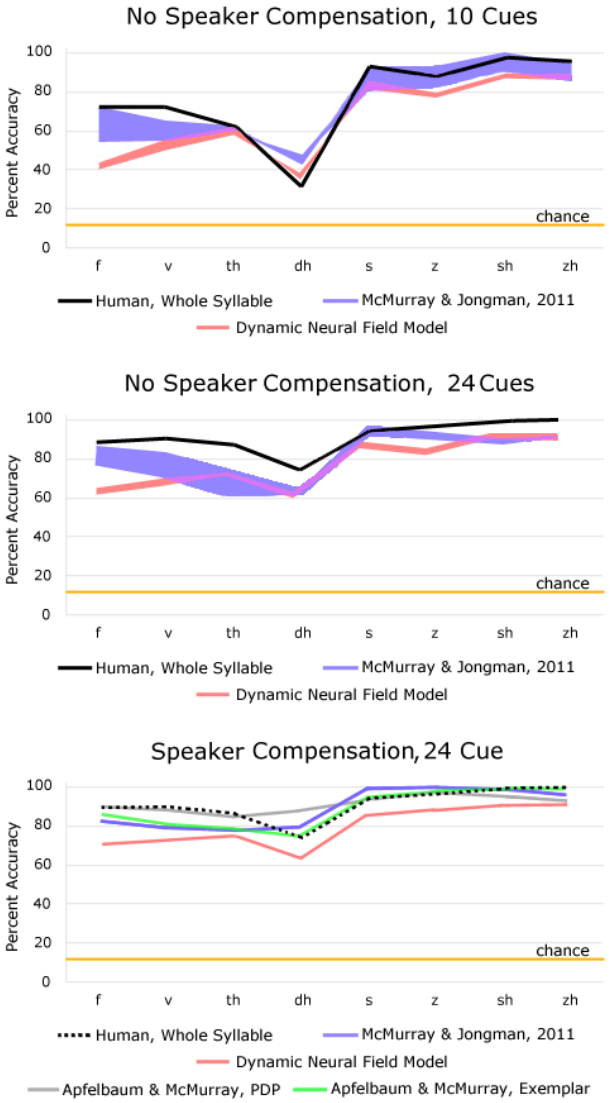

Figure 2. Simulation results. Shaded regions in the top two panels represent the accuracy range bounded by discrete-choice and probabilistic accuracy. The bottom panel shows only the mean of these two measures and dashed target line for easier reading.

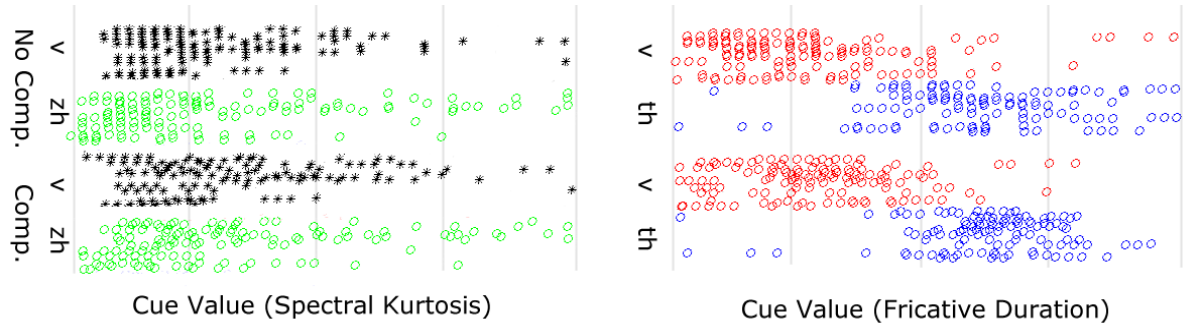

Figure 3. Cue values for test speakers before and after speaker compensation in the DNF model. Spectral kurtosis results in lower accuracy after compensation, while fricative duration results in higher accuracy.

Discussion

The DNF model establishes a set of plausible neural mechanisms for categorizing speech cues and compensating for variance between speakers. Accuracy was unsurprisingly lower than more abstract models, but only slightly so, and the qualitative pattern of DNF model results realistically follows that of human behavior.

The weakest portion of the DNF model quantitatively is its 'f' and 'v' accuracy. The possibility that some cues like spectral kurtosis were unhelpful for these phonemes when compensating for speaker was investigated and rejected as a hypothesis. An alternative explanation for this relative weakness is a high priority for future modeling.

Other directions for future work involve capitalizing on the rich representations that exist in a neural implementation and removing automated processes in model simulations.

The DNF model currently pre-loads some information into long term memory fields, like speaker adjustment profiles (as do competitor models). One high priority for model improvement is to remove this artificial seeding of information into the model and replace it with online learning of speaker adjustments. Speaker adjustments will be learned when context clues in the speech environment provide information about a speaker's intended phoneme beyond speech cues alone. Correct phoneme information can activate expected values for speech cues for that phoneme. These values can then be subtracted from the raw, perceived speech cue values of the speaker, using a transform field exactly like the one in the green region of Figure 1. The resulting adjustment value will then be stored for use later when context clues are unavailable (achieving the current starting point of the model).

The DNF model also currently blocks trials by speaker in order to conveniently activate one speaker memory profile at a time without rapid switching. Humans are able to utilize speaker information per syllable, however, so adapting the model to have this capability is a third modeling priority. This improvement requires only parameterization and gating of existing units to achieve more reliable timing.

Some of the DNF model's features suggest testable predictions for future investigation as well. For example, the DNF model naturally accounts for both prototype and exemplar memory representations, with no qualitatively different architecture than is described above. The dimension-coded LTM fields (right side of Figure 1) can store a profile of information about a specific speaker's history of one auditory cue's values, but could as easily store a profile for "males" or "females" in general, or for a specific moment of speech in time. Such stored information should be able to capture known human behaviors. For instance, Johnson, Strand, and D'Imperio (1999) presented discrimination results between speakers by gender that may be captured by DNF modeling. Artificial groups of phoneme tokens should also be able to be constructed grouped by a features beyond gender or speaker, such as visual scene context, arbitrary label, or otherwise. The DNF model would predict that active compensation for any such grouping may be feasible, useful, and actually utilized by humans in categorizing those phonemes.

The fact that the DNF model can represent 24-cue stimuli at all in a neurally plausible way is an advantage over some accounts of stimulus representation that will be explored further. McMurray and Jongman (2011) established the importance of considering many cues for accuracy. Many models have represented multidimensional stimuli, however, only in an abstract, n-dimensional "feature space" (Richardson, 1938; Nosofsky, 1986). Such a space cannot exist biologically, since a handful of dimensions requires more neurons than exist in the brain. The DNF model offers a solution to this problem: the architecture for a single cue (Figure 1) need only be replicated linearly for additional cues. 24-cue stimuli require only 24 times more neural resources than single cue stimuli.

This advantage of easily incorporating many new dimensions of information extends beyond auditory cues to dimensional information of nearly every other type. Visual dimensional features, for example, like size, orientation, spatial frequency, or color hue can be represented in much the same way as speech cues. Critically, just as voice features can be compensated and discounted, then, so can environmental lighting, distance, or angle of view. The current model can therefore potentially provide a unified explanation of effects like shape or size constancy as a

result of the same or a similar mechanism to phoneme constancy shown here. The model may similarly be relevant to aftereffects in "high-level" sensory processing, like face distortion adaptation (Köhler & Wallach, 1944), as a result of adjustment peaks requiring a short period of time to shift or die away when a stimulus is changed.

A final advantage of the DNF model is that it can process several feature or cue values at once through its compensation mechanism. Only one adjustment is used here for practical simulation time during initial modeling, but any number of parallel fields could be used with only linearly increasing neural investment, allowing many types of compensation at once. This is unlikely to help phoneme categorization further, but it would predict an advantage for compensating for context information in whole "scenes" of information in other modalities, such as distinguishing between "forest sounds" and "jungle sounds," benefitting from many parallel compensations along diverse dimensions at once. The DNF model not only predicts the capability of humans to perform *normalized* categorizations of this sort, but it predicts specific dynamics. If cue values are distinct, they should not interfere with one another, but if near one another along a dimension, lateral dynamics within the initial perceptual fields (blue region, Figure 1) fields should sharpen both peaks, or cause them to merge, etc., leading to distinct predicted categorization decisions.

Overall, compensation performance in the DNF model is promising. Future work will focus on increasing the model's self-sufficiency without pre-loaded information, investigation of novel behavioral predictions of the model, and expanding simulations to other sensory domains.

Acknowledgments

This research was supported in part by a Discovery Grant awarded by NSERC Canada and a Tier II Canada Research Chair. The authors would like to thank Bob McMurray and Allard Jongman for their prior model and consultation.

References

Apfelbaum, K., & McMurray, B. (2015). Relative cue encoding in the context of sophisticated models of categorization: Separating information from categorization. Psychon Bull Rev, *22*, 916-943.

Bendor, D. & Wang, X. (2005). The neuronal representation of pitch in primate auditory cortex, *Letters to Nature, 436*, 1161-1165.

Erlhagen, W. & S. & Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychological Review, 109*, 545-572.

Faubel, C. & Schöner, G. (2008). Learning to recognize objects on the fly: A neurally based dynamic field approach. *Neural Networks, 98* (3), 419-456.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psych. Rev., 105*, 251-279

Graves, A., Mohamed, A.-R., and Hinton, G. E. (2013). Speech recognition with deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 6645-6649.

Groh, J. M. (2001). Converting neural signals from place codes to rate codes. *Biological Cybernetics, 85*, 159-165.

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics, 27*, 359-384.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America, 108*, 1252-1263.

Köhler, W. & Wallach, H. (1944). Figural after-effects: An investigation of visual processes. *Proceedings of the American Philosophical Society, 88*, 269-357.

McMurray, B., Cole, J. S., & Munson, C. (2011). Features as an emergent product of computing perceptual cues relative to expectations. In C.N. Clements and R. Ridouane (Eds.), *Where do Phonological Features Come From? Cognitive, Physical, and Developmental Bases of Distinctive Speech Categories* (197-236). Amsterdan, John Benjamins.

McMurray, B. & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychogical Review, 2*, 219-246.

McMurray, B., Horst, J. S., Toscano, J. C., & Samuelson, L. K. (2009). Integrating connectionist learning and dynamical systems processing: Case studies in speech and lexical development. In Spencer, J. P., Thomas, S. C., & McClelland, J. L. (Eds.), *Toward a Unified Theory of Development* (218-249). Oxford University Press.

Miller, J. L. (2001). Mapping from acoustic signal to phonetic category: Internal category structure, context effects and speeded categorization, *Language and Cognitive Processes*, *16*(5/6), 683-690.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*(1), 39-57.

Pisoni, D. B. (1997). Some thoughts on "normalization" in speech perception. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (9-32). San Diego, CA: Academic Press.

Richardson, M.W. (1938). Multidimensional psychophysics. *Psychological Bulletin, 35*, 659-660.

Rock, I. (1983). *The logic of perception.* Cambridge, MA: MIT Press.

Samuelson, L. K., Spencer, J. P., & Jenkins, G. W. (2013). A dynamic neural field model of word learning. In Gogate & G. Hollich (Eds.), *Theoretical and computational models of word learning: Trends in psychology and artificial intelligence.* Hershey, PA, US: Information Science Reference / IGI Global.

Schneegans, S. & Schöner, G. (2012). A neural mechanism for coordinate transformation predicts pre-saccadic remapping. *Biological Cybernetics, 106*, 89-109.

Stevens, K. N. & Keyser, S. J. (2010). Quantal theory, enhancement and overlap. *Journal of Phonetics, 38*,10-19.