Brain Feedback and Adaptive Resonance in Speech Perception

https://hdl.handle.net/2144/2239

Boston University
BBS COMMENTARY
on "Merging information in speech recognition:
Feedback is never necessary"
by D. Norris, J.M. McQueen, and A. Cutler

Brain feedback and adaptive resonance
in speech perception

Commentator: Stephen Grossberg

July, 1999

Technical Report CAS/CNS-99-022
BRAIN FEEDBACK AND ADAPTIVE RESONANCE IN SPEECH PERCEPTION

Commentary for
Behavioral and Brain Sciences
on
Merging Information in Speech Recognition:
Feedback is Never Necessary
by
D. Norris, J.M. McQueen, and A. Cutler

Commentator: Stephen Grossberg
July 21, 1999

Boston, MA: Boston University

The brain contains ubiquitous reciprocal bottom-up and top-down intercortical and thalamocortical pathways. These resonating feedback pathways are suggested to be essential for stable learning of speech and language codes, and for enabling context-sensitive selection and completion of noisy speech sounds and word groupings to occur. Context-sensitive speech data, notably data involving interword backward effects in time, have been quantitatively modeled using these concepts, but have not been modeled by purely feedforward models.

Norris et al. argue that “top-down feedback does not benefit speech recognition” and that “no experimental data imply that feedback loops are required for speech recognition. Feedback is accordingly unnecessary.” They carry this position perhaps as far as it can go, and nicely describe how their feedforward MERGE model can explain some data at least as well as the feedback TRACE model and the feedforward RACE model. They focus on TRACE as a representative feedback model because it is “the main standard bearer of interaction.” This is a debatable assumption because TRACE has major conceptual and data-predictive problems that are not shared by other feedback models (Grossberg, Boardman, and Cohen, 1997). On the conceptual side, TRACE is not a real-time model, cannot self-organize, and experiences a major combinatorial explosion. On the data side, TRACE cannot explain a host of data in which backward effects contextually alter speech percepts. FLMP also has such problems. Norris et al. are also selective in their choice of psychological and neural data with which to support their thesis, and underplay serious conceptual problems with their own model that feedback models have already overcome.

Massive and selective feedback processes exist in every cortical and thalamic region (Felleman and Van Essen, 1991). Norris et al. are claiming that these processes play no role in speech recognition. In fact, neural models have recently suggested how the laminar circuits of neocortex merge feedforward, horizontal, and feedback pathways to elegantly achieve three goals: (1) stable development of cortical connections and adult learning; (2) seamless fusion of bottom-up and top-down processing, whereby top-down feedback modulates, matches, and attentively selects bottom-up data that are consistent with learned top-down hypotheses; and (3) a synthesis of analog representation and coherent binding of distributed information that is called analog coherence (Grossberg, 1999a; Grossberg, Mingolla, and Ross, 1997).

Norris et al. do not explain how a feedforward model can explain classical phonemic restoration data: Let a listener hear a broad-band noise followed rapidly by the words “eel is on the....” If this
word string is followed by “orange”, then “noise-eel” sounds like “peel”; if by “wagon”, it sounds like “wheel”; if by “shoe”, it sounds like “heel” (Warren, 1984; Warren and Sherman, 1974). If some formants of the expected sound are missing from the noise, then only a partial reconstruction is heard (Samuel, 1981a, 1981b). If silence replaces the noise, then only silence is heard, and the sentence meaning changes, e.g., consider “eel is on the shoe”. These results strongly argue that the feedforward signal is not what is consciously heard. Instead, contextual feedback from the meaning of the entire sentence “feeds backwards in time” across several words to select those noise formants that are consistent with a contextually sensitive top-down expectation. This top-down matching process cannot, however, “create something out of nothing. It can only select and focus attention on what is already in the feedforward data stream. This attentive process can take from 100 to 200 msec. to generate a conscious percept. It demonstrates an intimate interaction between lexical and prelexical processes.

Adaptive Resonance Theory (ART) models explain such data as properties of brain resonances that focus attention upon important bottom-up data while stabilizing brain development and learning (e.g., Boardman, Cohen, and Grossberg, 1997; Cohen and Grossberg, 1986; Grossberg, 1978, 1980, 1986, 1995, 1999b; Grossberg and Stone, 1986). The time scale of conscious speech is identified with the time needed for interacting bottom-up and top-down processes to achieve resonance. The matching properties help to stabilize brain development and learning.

There are many other examples of backward effects in time. Repp (1980) studied categorical perception of VC-CV syllables. He varied the silence interval between the VC and CV syllables in [ib]-[ga] and [ib]-[ba]. If the silence is short enough, then [ib]-[ga] sounds like [iga] and [ib]-[ba] sounds like [iba]. Remarkably, the transition from [iba] to [ib]-[ba] occurs after 100-150 msec more silence than the transition from [iga] to [ib]-[ga]. This is a very long interval for a feedforward model to bridge. Moreover, whether fusion or separation occurs at a given silence interval is context-sensitive. These data have been quantitatively explained by resonant fusion in the case of [iba] and resonant reset in the case of [iga] (Grossberg, Boardman, and Cohen, 1997). They illustrate the ART hypotheses that “conscious speech is a resonant wave” and that “silence is a temporal discontinuity in the rate with which resonance evolves”.

Repp et al. (1978) varied the silence interval between the words GRAY CHIP and the fricative noise duration in CH. They hereby generated percepts of GREAT CHIP, GRAY SHIP, and GREAT SHIP. Remarkably, increasing silence duration transforms GRAY CHIP into a percept of GREAT CHIP, and increasing noise duration can transform it into a percept of GREAT SHIP. Why should more silence or more noise in a future word convert a past word GRAY into GREAT? Why should more noise remove the CH from CHIP and attach it to GRAY to form GREAT, leaping over a silent interval to do so, and becoming detached from its contiguous word? These effects have also been quantitatively simulated by ART (Grossberg and Myers, 1999).

The MERGE model shares some key processes with ART, such as competition between activated lexical hypotheses, multiple interactive activation cycles, and reset events (Grossberg, 1980; Grossberg and Stone, 1986). But MERGE also has serious weaknesses due to its feedforward structure. It keeps lexical and prelexical computations independent until they are merged at the decision stage. How this scheme can naturally explain the backwards-in-time data above is unclear. MERGE’s feedforward decision stage is, moreover, not a real-time physical model: “the word nodes cannot be permanently connected to the decision nodes...the connections...must be built on the fly, when the listener is required to make phonemic decisions...decision nodes...set up in response to a particular experimental situation.” This cannot be how the brain works. In addition, the MERGE decision stage represents both phonemic and lexical information in a way that can “translate the presentations used for lexical access into the representations more suited to...phonemic decision tasks.” How and why this should happen is left unclear.
ART naturally overcomes these problems using evolving spatial patterns of activation across working memory items that resonate with a level of list chunks. The list chunks that are learned in this way can include phonemic, syllabic, and word representations. The resonant context determines which chunks are competitively selected and learned. A Masking Field architecture was introduced to represent list chunks of variable length. It explains how phonemic, syllabic, and lexical information can coexist at the list chunk level, and how the speech context determines whether phonemic or lexical information will dominate (Cohen and Grossberg, 1986; Grossberg, 1978). Thus, there is no need to generate connections on the fly. This property helps to explain the Magic Number 7, word length and superiority effects, the GRAY CHIP percepts, and why phonemic decisions may not develop prior to word recognition, among other data.

In summary, the feedforward MERGE model has not yet solved core problems for which the feedback ART model has proposed real-time, neurally-supported, self-organizing, and data-predictive solutions.
REFERENCES


