NEURAL DYNAMICS OF PERCEPTUAL ORDER AND

CONTEXT EFFECTS FOR VARIABLE-RATE SPEECH SYLLABLES

by

Ian Boardman[1], Stephen Grossberg[2], Christopher Myers[3], and Michael Cohen[4]

Department of Cognitive and Neural Systems
and
Center for Adaptive Systems
Boston University
677 Beacon Street
Boston, MA 02215

Suggested Running Head: Neural Dynamics of Speech Context Effects

Please address reprint inquiries to:
Stephen Grossberg
Department of Cognitive and Neural Systems
Boston University
677 Beacon Street
Boston, MA 02215

Technical Report CAS/CNS-TR-98-004
Boston, MA: Boston University.

ABSTRACT

How does the brain extract invariant properties of variable-rate speech? A neural model, called PHONET, is developed to explain aspects of this process and, along the way, data about perceptual context effects. For example, in consonant vowel (CV) syllables such as /ba/ and /wa/, an increase in the duration of the vowel can cause a switch in the percept of the preceding consonant from /w/ to /b/ (Miller and Liberman, 1979). The frequency extent of the initial formant transitions of fixed duration also influences the percept (Schwab, Sawusch, and Nusbaum, 1981). PHONET quantitatively simulates over 98% of the variance in these data using a single set of parameters. The model also qualitatively explains many data about other perceptual context effects. In the model, C and V inputs are filtered by parallel auditory streams that respond preferentially to transient and sustained properties of the acoustic signal before being stored in parallel working memories. A lateral inhibitory network of onset- and rate-sensitive cells in the transient channel extracts measures of frequency transition rate and extent. Greater activation of the transient stream can increase the processing rate in the sustained stream via a cross-stream automatic gain control interaction. The stored activities across these gain-controlled working memories provide a basis for rate-invariant perception, since the transient-to-sustained gain control tends to preserve the *relative* activities across the transient and sustained working memories as speech rate changes. Comparisons with alternative models tested suggest the fit can not be attributed to the simplicity of the data. Brain analogs of model cell types are described.

Key words : context effects, CV syllable, formant transition, neural network, phonetic perception, speech perception, vowel, consonant, sustained cells, transient cells, transition duration, transition rate, working memory.

1. Introduction

A challenging problem in cognitive psychology and neuroscience is to develop a predictive theory of how humans understand language when it is spoken at different rates. An important aspect of this problem involves understanding how the listener integrates acoustic segments that vary with speech rate into unitary percepts of consonants and vowels and further integrates these into words. The search for rate-invariant acoustic features has not met with notable success (Assman, Nearey, and Hogan, 1982; Bailey and Summerfield, 1980). Part of the difficulty stems from the fact that, as Repp and Liberman (1987) put it, "phonetic categories are flexible": a given acoustic segment can be perceived as one phoneme or another depending on its context. For example, effects can occur in which varying the duration of the subsequent vowel /a/ can alter the percept of the preceding consonant from /b/ to /w/. This leads to either a /ba/ or a /wa/ percept (Miller and Liberman, 1979). The present article develops a neural model, called PHONET, which suggests an explanation of how such context effects can occur. PHONET simulates these effects using model neural mechanisms whose functional role is to transform rate-varying properties of consonant-vowel transitions and steady-state vowels into an internal representation from which rate-invariant properties can be extracted.

Several problems need to be analyzed to understand these context effects and our model of them: How does the brain represent consonant and vowel features? How are both sorts of speech sounds temporarily stored in a working memory so that a subsequent event, such as a change in vowel duration, can alter the percept of a preceding consonant before it reaches conscious awareness? How do the working memory representations of consonants and vowels interact? Why does the conscious percept take so long to emerge that the duration of the subsequent vowel *can* influence the percept of a preceding consonant? Why do not listeners already consciously perceive the consonant before the vowel is fully presented? Finally, how do these several processes interact to ensure that language can be understood even if it is spoken at different rates?

We address these questions by integrating psycholinguistic and neural data in a theoretical framework which has already been used to successfully describe related speech perception data (Bradski, Carpenter, and Grossberg, 1994; Cohen and Grossberg, 1986, 1997; Cohen, Grossberg, and Stork, 1988; Grossberg, 1978, 1986, 1995, 1998; Grossberg, Boardman, and Cohen, 1997; Grossberg and Stone, 1986a, 1986b). Section 2 describes some of the empirical issues associated with variable-rate speech perception and the concept of working memory storage of acoustic tokens in the formation of phonetic percepts. Section 3 provides a more detailed overview of the type of context effects that will be analyzed. The experiments of Miller and Liberman (1979) and Schwab, Sawusch, and Nusbaum (1981) are examined in particular. Section 4 and thereafter shows how PHONET quantitatively simulates the Miller and Liberman (1979) and Schwab *et al.* (1981) data and qualitatively explains other related context effects.

2. Temporal Variation in Speech

Variations in the rhythm and tempo of speech are a natural part of human communication. Our everyday experience provides examples of the influence of timing on meaning.

Timing cues are important in distinguishing word and syllable junctures (e.g., topic and top pick), voicing contrasts (e.g., /b/ and /p/), and manner distinctions (e.g., /sh/ and /ch/). These examples of moment-by-moment or *local* changes in the speech rate cue perception of different phonemes and signal distinctions in meaning. We are also able to interpret the speaker's meaning even with variation of the long-term or *global* speech rate. Speech understanding remains invariant even while the durations of speech sounds scale, often in complex ways, with the duration of a phrase or larger speech frame (see Miller, 1981).

Speech tempo is a reflection of the durations of acoustic segments characterized by their distribution of energy across frequency; i.e., their spectral patterns. Speech segments can be grossly described as intervals of either rapidly changing frequency content or of quasi-static frequency content, and they have spectral features that are highly correlated to consonant and vowel productions, respectively. While this distinction between "transient" and "sustained" spectra fails to accurately describe the detailed spectral patterns of all consonants and vowels in all contexts, it serves as a useful first pass in identifying their significant acoustic correlates. Psychoacoustic evidence, described below, shows that our auditory system distinguishes these two types of segments and is sensitive to their durations. Data from auditory physiology also support the view of distinct neural mechanisms for coding of transient sounds versus sustained sounds, such as steady-state frequencies, which are periodic in nature (e.g., Phillips, 1993; Pickles, 1988).

Some phonemes have clear acoustic correlates; for example, vowels are often characterized by steady acoustic energy concentrated in several narrow bands in frequency space called *formant* frequencies. However, acoustic segments do not necessarily correspond directly to phonetic percepts. For example, most consonants cannot be perceived as phonemes independent of the vowel context. The vowel sound that precedes or follows the production of the consonant is perceived as a vowel, while simultaneously contributing to the percept of the consonant. In addition, the articulatory maneuvers associated with phonemes are often distributed in time and interleaved with the articulations of past or future speech sounds, a phenomenon called *coarticulation*. The acoustic consequences of these coarticulated gestures are similarly distributed and interleaved in time, making it necessary to integrate or *group* speech sounds across time to derive unified phonetic percepts.

Grouping of acoustic features operates on the local time scale of acoustic segments, which can be considered roughly 10–100 msec although these limits vary with speaking rate and other factors. This featural grouping is sensitive to temporal information *intrinsic* to the segments such as segment durations (Repp, 1978; Miller, 1987). The acoustic features that are grouped are represented by spatial patterns of activation over tonotopically organized fields in the auditory system (Pickles, 1988). Grouping suggests a compression of these broadly distributed spatial patterns into more narrowly focussed activations, or item representations, at a processing stage that supports context–sensitive phonetic percepts.

Recognition of meaningful language units, such as words and phrases, requires another stage of grouping that is sensitive to serially ordered lists of item representations that are stored in a working memory (Baddeley, 1986; Bradski, *et al.*, 1992, 1994; Grossberg, 1978; Grossberg, Boardman, and Cohen, 1997; Grossberg and Stone, 1986a). Lists of active item representations in working memory are compressed into activations of list chunks at a later processing stage. Grouping of items into list chunks operates on the global time scale of syllables and is sensitive to *extrinsic* temporal information; e.g., the global speech rate (Repp,

1978; Miller, 1987).

The focus of this article concerns the process by which phonetic percepts depend upon temporal information present in the local context. The research addresses how current speech events can influence the percepts associated with prior speech events, as when the identification of a syllable-initial consonant varies according to the duration of the following vowel (Miller and Liberman, 1979). In the PHONET model that is developed here, transient and sustained components of the acoustic features interact in a duration-dependent manner. These interactions generate a rate-compensating activation pattern of item representations in working memory that provides an invariant basis for subsequent phonetic recognition. In particular, the interactions cause the ratio of activities across the transient and sustained working memories to remain invariant as speech rate changes and thus represent a stable phonetic code. This activity ratio also exhibits observed context effects, such as how the duration of a subsequent vowel can influence the percept of a preceding consonant. In other words, mechanisms aimed at ensuring rate-independent speech perception can also explain rate-dependent context effects.

3. Context Effects in CV Perception

Denes (1955) varied the duration of acoustic segments perceived as consonants and vowels to study the voicing differences between the pronunciations of the word "use". Strictly speaking, "consonant" and "vowel" refer to phonetic, not acoustic, representations. However, for simplicity when discussing an acoustic signal perceived as a CV syllable, we will refer to the duration of formant transitions as the *consonant duration* and the duration of the steady-state formants as the *vowel duration*. Denes concluded that the ratio of consonant to vowel duration is a cue to voicing of the final consonant, such that perceived voicing increases with the C/V ratio.

The early experiments of Denes (1955) represent one particular class of phonetic context effects that has been termed *durational contrast* (Diehl and Walsh, 1989). The primary acoustic distinction between some consonant pairs lies in their duration. For example, the spectra of the fricative /sh/ and the affricate /ch/ tend to be very similar, but in natural speech /sh/ sounds are typically 60 msec longer than /ch/ sounds (Howell and Rosen, 1983). Similarly, the formant transitions that distinguish the perceived manner between the stop /b/ and the semi-vowel /w/ differ primarily in duration, and secondarily in other qualities such as rise-time of the amplitude envelope (Liberman, *et al.*, 1956; Shinn and Blumstein, 1984). When two phonemes differ primarily in duration, their category boundary can be shifted by varying the duration of adjacent segments. In general, such dependencies demonstrate a durational contrast: "an interval will be heard as shorter in the context of a long segment than in the context of a short segment" (Diehl and Walsh, 1989). Durational contrast operates not only in manner distinctions such as /sh/-/ch/ and /b/-/w/, but in voiced-voiceless (e.g., /b/-/p/) and single-geminate (e.g., [iba]-[ibba]) distinctions. Two accounts of durational contrast have been given, one based on speaking rate normalization (e.g., Miller, 1981), the other on principles of general auditory processing (Diehl and Walsh, 1989; Pisoni, Carrell, and Gans, 1983).

Studies by Miller and Liberman (1979) and Schwab *et al.* (1981) provide a source of careful parametric data that help to reveal the structure of some durational context effects.
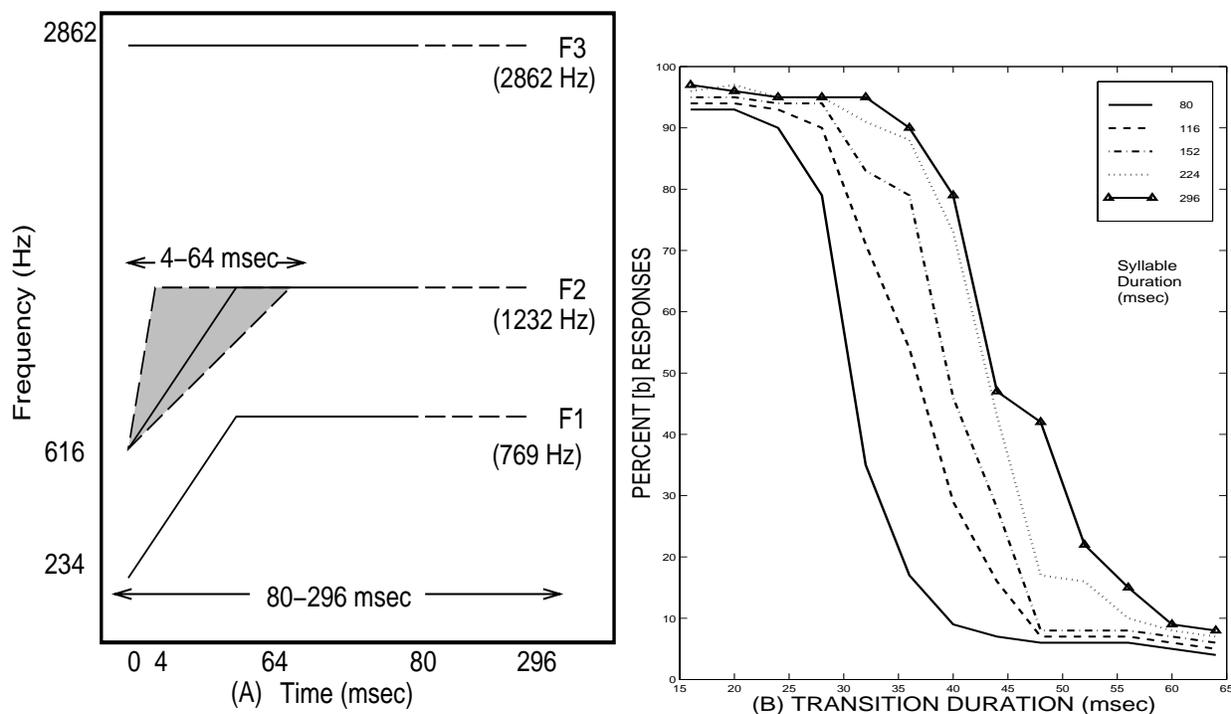
Figure 1: (A) Schematic diagram of CV stimuli used by Miller and Liberman (1979), indicating formant frequencies for steady-state (vowel) and transition (consonant) segments. Segment durations were varied. (B) Original data of Miller and Liberman (1979), reproduced with permission from their Figure 3. Each curve describes the probability of [b] percept as a function of transition duration, indicated along the abscissa, for a fixed syllable duration, indicated below the curve.

Miller and Liberman (1979) presented subjects with artificially generated CV syllables that were identified either as /ba/ or /wa/, depending on the durations of the consonant and vowel segments. Their speech stimuli had the stereotyped spectral patterns depicted in Figure 1A. The formant transitions increased linearly from fixed starting frequencies and ended at the corresponding vowel (steady-state) formant frequencies. Four sets of stimuli were created, each with a different fixed vowel duration ranging from 80 to 296 msec. In each set the duration of the formant transitions varied from 16 to 64 msec in increments of 4 msec. The findings of the study, summarized in Figure 1B, showed that the following vowel duration influences the percept of the preceding consonant, as reflected in the subjects' probability of responding /b/. The rightward shift of the boundaries with syllable duration implies that, with transition duration fixed, the probability of /b/ increases with vowel duration. On the other hand, the percept may remain unchanged (probabilistically) with longer transition intervals if the following vowel durations are also longer by appropriate amounts. Summerfield (1981) reported data consistent with this finding, among other results. Pisoni, *et al.* (1983) described analogous context effects using non-speech analogues. They used four or five sinusoid tones, each of which represented the middle frequency of one formant. Each sinusoid tone was crafted to follow the frequency contour of the formant, including its onset transient, steady state, and offset transient.
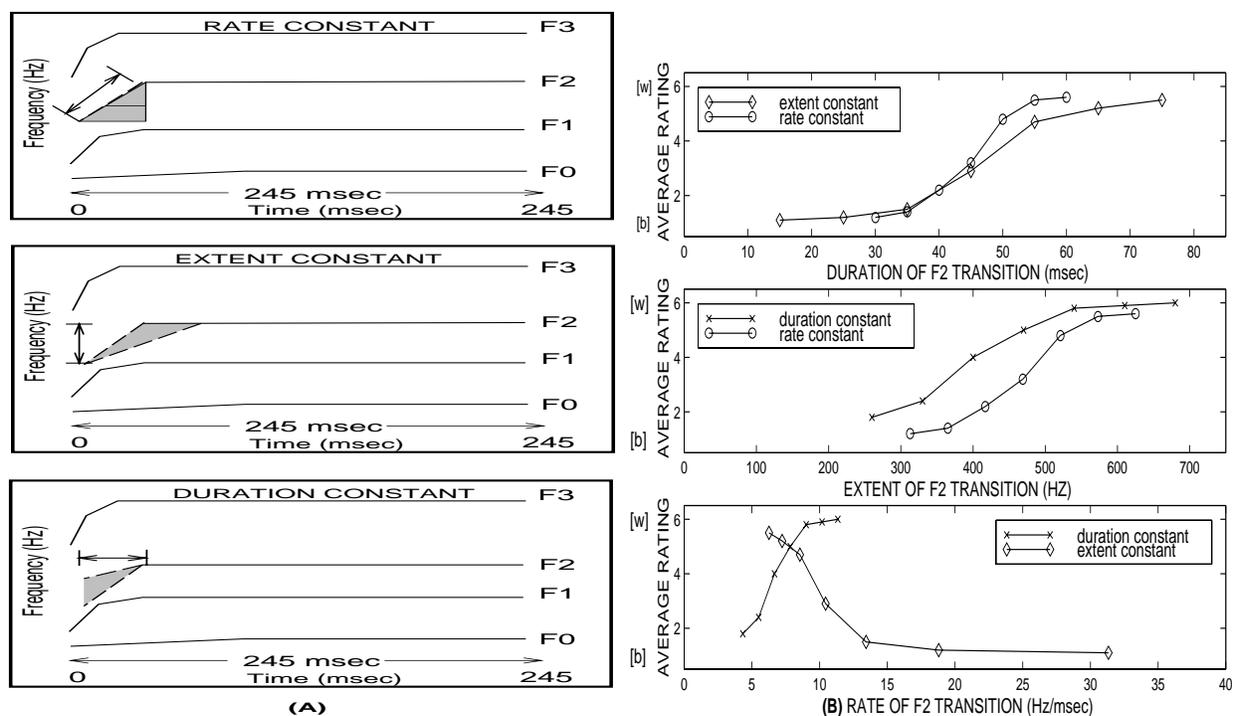
Figure 2: (A) Schematic diagram of CV stimuli used by Schwab, Sawusch, and Nusbaum (1981), redrawn from their Figures 1-3. In each series, one of the duration, rate, or frequency extent of the F2 formant transition is fixed while the other two vary across seven values. (B) Original data of Schwab, Sawusch, and Nusbaum (1981), redrawn from their Figures 4-6. Stimuli were rated on a scale of 1 ("good /ba/") to 6 ("good /wa/"). Curves show average rating functions for the pairs of series that vary F2 transition duration, extent, and rate.

Schwab, Sawusch, and Nusbaum (1981) further studied which combinations of acoustic cues determine whether the stop consonant /b/ or semivowel /w/ is the percept obtained in /ba/ or /wa/, respectively. In the Miller and Liberman (1979) experiments, the transition *duration* leading to a /b/ or /w/ percept always covaried with the transition *rate*, with the frequency extent of the transition kept constant (Figure 1A). Therefore, transition duration and rate varied together across trials. Schwab *et al.* (1981) independently varied these transition properties. They concluded that transition duration (D) and frequency extent (E) – namely, the total frequency change – but not rate, specify the /b/-/w/ distinction. In particular, they noted that if the product E · D exceeded a criterion of 23,000 Hz · msec for their subjects, then /w/ was perceived, whereas if E · D was less than this criterion, then /b/ was perceived; this rule held in 196 of 210 judgments.

Schwab *et al.* (1981) ran three experimental series. In the first, shown schematically in Figure 2A, they fixed the F2 transition rate at 10.43 Hz/msec; in the second (Figure 2B), they fixed the F2 transition extent to be 417 Hz; and in the third (Figure 2C), they fixed the F2 transition duration to be 60 msec. Since Duration × Rate = Extent, fixing any two of the parameters determines the third. At the fixed values of duration, extent and rate used in this experiment, Schwab *et al.* (1981) showed that the long duration of formant transition was a positive cue for /wa/, as was a large frequency extent of the formant transitions.

A fast rate of transition cued either /ba/ or /wa/ according to whether the extent or the duration was held constant. At constant extent, a high transition rate cues /ba/. However, at constant duration, the increasing rate cues /wa/.

Because Schwab *et al.* (1981) fixed the total duration of their syllables, their experiments did not probe effects depending on syllable length. The experiments of Miller and Liberman (1979) and Schwab *et al.* (1981) are thus complementary. Contrary to Schwab *et al.*'s (1981) proposal that transition extent and duration determine the percept, Miller and Liberman (1979) showed an effect of total syllable length, by varying duration. When syllable length is fixed, Schwab *et al.* (1981) showed an effect of the frequency extent of the formant transitions.

Shinn and Blumstein (1984) established that, in addition to the spectral characteristics of the formant transitions, the amplitude envelope of the consonantal release also helps determine the perceived identity of /ba/ or /wa/. In their experiments, two series of synthetic /ba/-/wa/ stimuli were constructed. Each stimulus in one series was given a stop envelope, and each stimulus in the other series was given a glide envelope. These cues appeared to strongly override the cues carried by formant transition rate and duration. However, Nittrouer and Studdert-Kennedy (1986) obtained the opposite result using natural stimuli. When amplitude envelopes were exchanged, 97% of the stimuli were identified as originally produced. These two experiments imply that, while amplitude envelopes can influence the /ba/-/wa/ distinction in synthetic stimuli, it appears that natural and synthetic stimuli are not directly comparable.

Walsh and Diehl (1991) showed that neither extent nor duration of formant transitions, nor subsequent vowel length, totally predicts the /ba/ or /wa/ percept. The rise time of the amplitude of the second formant is an independent cue for /ba/. However, they also showed that this is a less potent cue for the /ba/–/wa/ distinction than either syllable duration or formant transitions. A later study by Shinn *et al.* (1985) showed that the size of the syllable duration effect was greatly reduced, or even eliminated, if natural speech was used. Miller and Wayland (1993) found, however, that under noisy conditions, which may arguably be more typical during everyday speech, even the stimuli used by Shinn *et al.* (1985) lead to duration-sensitive context effects. Derr and Massaro (1980) and Port and Dalby (1982) further studied the effects of duration on CV perception. Port and Dalby (1982) concluded, with Denes (1955), that the CV ratio is a cue to voicing. Massaro and Cohen (1993) showed that both consonant duration and vowel duration, not just their ratio, may be used to provide a better fit to categorical data.

Newman and Sawusch (1996) recently studied several factors' influence on durational contrast, including similarity, phonotactics, and duration of the phonemes following a target phoneme. The authors extended the results of Miller and Liberman (1979) to show the effect of adjacent vowel duration on different initial phonetic contrasts (/ch/ vs. /sh/). They also found that stop consonants, much briefer than vowels, could induce a small but significant shift in the perception of initial phonetic contrasts (/ch/ vs. /sh/ and /sw/ vs. /tw/), even with the stop consonants separated from the target by an intervening vowel. The determining factor in their studies appeared to be a temporal window of about 300 msec following the target phoneme, over which perceived rate of articulation is integrated to influence identification of the target. This view is consistent with the model developed below, in which an exponential saturation of activity in the sustained channel limits the

temporal window over which the following vowel can affect the representation of an initial segment via the sustained/transient ratio.

Another line of experiments related to the present analysis concerns asymmetric vocalic context effects. Kunisaki and Fujisaki (1977) examined how vocalic (V) information and frication (F) information contribute to fricative perception. They used synthetic fricative noises that varied in spectral peak frequency whose percepts ranged from /ʃ/ to /s/ when presented alone. These noises were then embedded in synthetic VF and FV syllables. The authors observed that in an FV syllable a *following* vowel (/u/ or /o/) shifts the perceptual boundary to produce more /s/ responses, but in a VF syllable a *preceding* vowel does not affect the boundary.

Mann and Repp (1980) replicated these findings with FV and VF syllables using naturally produced vocalic segments. Mann and Soli (1991) followed up the Mann and Repp (1980) study with a number of additional controls. They again reported an asymmetric vocalic context effect, with FV pairs causing a much larger context effect than the VF pairs. Additional manipulations included playing stimuli backward to dissociate segmental order in the listening tasks from the order in which the segments were produced. The authors concluded that "the order of the vocalic and frication segments in the listening task, not the order in which they were originally produced, emerged as the primary determinant of the asymmetric vocalic context effects" (p. 409), and that "no single model can adequately account" for these effects (p. 410).

The asymmetric vocalic context effects call attention to the question: what differences between consonant and vowel processing are the basis for this asymmetry? The PHONET model suggests such a difference by which processing of consonantal transitions influences subsequent vowel processing. PHONET shows how such asymmetry can transform rate-variant CV acoustic signals into an internal representation in working memory from which rate-invariant phonetic information can be extracted.

All of the above experiments provide information pertinent to the development of our model. While our model qualitatively addresses many of their findings, we restricted our fit to the data of Schwab *et al.* (1981) and Miller and Liberman (1979). Our successful fits do not detract from the conclusion that more experiments are needed in this area, for at least three reasons.

First, in both the experiments of Miller and Liberman (1979) and Schwab *et al.* (1981), the psychometric curves obtained are averaged over all of the subjects. Lacking tests to determine homogeneity of the subject population, one cannot be sure that the average curves closely reflect the individual subjects' behavior. Second, the experimental stimuli in the Miller and Liberman (1979) and Schwab *et al.* (1979) studies differ in several respects. Schwab *et al.* (1981) varied formant F2 only, whereas Miller and Liberman (1979) varied both F1 and F2. Additionally, the Schwab *et al.* (1981) stimuli contained five formants, while the Miller and Liberman (1979) stimuli contained only three. Third, the Schwab *et al.* (1981) experiments consisted of rating tasks, while the Miller and Liberman (1979) data came from two alternative forced choice tasks. More extensive experimentation is necessary to compare the results of the two methods in these tasks.

4. PHONET Model: Sustained and Transient Processing Streams

Speech contains acoustical energy that exhibits both time-varying (transient) and steady-state (sustained) properties. For example, the CV stimuli used by Miller and Liberman (1979) contain energy concentrated in narrow regions across frequency space as formants, which are initially in motion and then attain a stable position. The brain can represent these acoustical features as time-varying spatial patterns of neural activity. The representation begins with the sensory tissue that transduces acoustical signals into neural activity and maps frequency into the spatial dimension along the auditory nerve fiber array. This mapping is called a *tonotopy*, and it is preserved from the organ of Corti in the inner ear through the thalamus and on up to the auditory cortex (Irvine, 1986; Pickles, 1988). One finds a similar configuration in vision, where regions of luminous energy projected on the surface of the retina are mapped into spatial patterns of neural activity across the surface of the primary visual cortex. Movement of features in the visual scene is detected at the retina, and transient motion information is transferred separately from, but in parallel with, the more sustained form information on up through the thalamus to distinct areas of visual cortex (DeYoe and Van Essen, 1988).

Various data suggest that the auditory system also is separately sensitive to transient and sustained information in the "auditory scene." Cells in the cochlear nuclei of rats and cats are usefully classified by their temporal response to brief tones near their best frequency (Pickles, 1988). Some of these cells show prolonged responses through the duration of the tone, while others respond primarily to onsets of acoustical energy in a particular frequency band (Pickles, 1988; Rhode and Smith, 1986). In principle, these cells could signal onsets of vocal pitch pulses or passage of a formant peak through their region of frequency selectivity. Additionally, certain cell types show selectivity to linear FM sweeps of a particular rate in cat cochlear nucleus (e.g., Britt and Starr, 1976b) and cat auditory cortex (e.g., Mendelson, *et al.*, 1993; Tian and Raushecker, 1994). Such linear FM sweeps provide good approximations to speech formant transitions. Cells were discovered in rat cochlear nucleus that did not respond significantly to tones, but did respond when the stimulus frequency was ramped up and down at a particular rate (Møller, 1983). The tuning to sweep rate appears to be topographically organized in cat auditory cortex. Using multiunit recording, Mendelson *et al.* (1993) mapped out responses of primary auditory cortex (A1) to linear FM sweeps of varying speeds. In a dorsal to ventral progression, cells were selective to fast FM rates, then medium, slow, medium, and, most ventrally, fast rates. Mendelson *et al.* (1993) concluded that "the overall distribution of these response parameters within A1 suggests that in addition to tonotopicity, the functional organization of A1 may also be defined by an orthogonal dimension of FM sweep responses within the isofrequency domain" (Mendelson *et al.*, 1993, pp. 80-81). These rate-sensitive cells show transient activation as the sweep passes through their best frequency band, behaving like neurons in rat inferior colliculus in that they code "the transient character of a FM signal, the timing of how fast it enters and leaves the excitatory response area of a neuron" (Felsheim and Ostwald, 1996, p. 150).

Sounds which contain energy in a constant frequency region for a relatively sustained length of time, like pure tones or steady-state vowels, may be processed by cells that show prolonged responses over the course of the stimulus. Properties of the eighth nerve exemplify the ability of the auditory system to selectively process sustained vowel-like sounds (Delgutte

and Kiang, 1984a, 1984b; Sachs and Young, 1979; Young and Sachs, 1979). Distinctive cell types of the cochlear nucleus known as primary, pauser, and buildup cells, show activity patterns that have characteristic temporal profiles extending throughout a pure tone stimulus (Britt and Starr, 1976a; Pickles, 1988; Rhode and Smith, 1986).

Cohen and Grossberg (1997) developed the hypothesis that the auditory system branches into parallel streams that preferentially process transient, or rapidly changing, and sustained, or steady-state, features of the acoustic signal into a neural model of early auditory filtering. The model's transient and sustained detectors begin to separate coarticulated consonants and vowels. Building upon this work, the present model makes use of neurons responsive to changes of auditory nerve fiber activity within a localized region of the nerve fiber array, corresponding to some narrow frequency band. These model neurons, which play the role of transient detectors, respond to the time rate of change in their inputs. They are sensitive to onsets, bursts and formant transitions with spectral energy in their receptive field. These model neurons interact within a recurrent on-center off-surround network, and thereby become sensitive also to the frequency extent of the varying stimuli.

Complementary model neurons act as sustained detectors, sensitive to the level of stationary or steady-state activity on the auditory nerve within a localized region. These respond to vowels and nasals with spectral energy in their receptive field. Both the transient and the sustained channels perform temporal and spatial integration of the respective type of detector signals that arise within the channel's receptive field. Averaging for the transient channel produces a response that is sensitive to both the rate of change of spectral features within the channel's frequency range and the frequency extent of the transient stimuli. Averaging the sustained channel produces a response sensitive only to slowly moving or stable spectral features within the channel's input space.

Other researchers have also suggested that functional analogies exist between cortical processing of speech and vision. For example, Shamma, Versnel, and Kowalski (1995) recorded cell responses in ferret primary auditory cortex (A1) to spectrally shaped noise, finding that FM direction selectivity is mapped along isofrequency planes in A1. Shamma and colleagues note the similarity between these "spectral gradient maps" and the maps of orientation selectivity known to exist in visual cortex. The responses of ferret A1 cells to spectral grating stimuli are analogous to visual cortical cells which have transfer functions "tuned around a specific grating frequency (usually called 'spatial frequency')" whose inverse transforms "predict well the receptive field of the cell measured by impulse-like stimuli as light dots" (Shamma *et al.*, 1995, p. 253). Furthermore, "it is crucial to recognize that apart from the dimensionality of the input signal, the mechanisms giving rise to orientation selectivity in V1 are identical to those seen in A1" (Shamma *et al.*, 1995, p. 253). Thus, Shamma *et al.* relate sensitivity to spectral motion in auditory cortex to edge orientation maps in visual cortex.

Others have endorsed the notion of parallel sustained and transient processing streams in audition. Arguing from temporal masking level difference experiments, Berg (1985) proposed that separate short-term and long-term integrators analyze interaural differences. According to whether the systems provide confirmatory or discrepant spatial location estimates, their output is combined or processed independently. Such a system explains her data showing "that increases in backward-masker duration produced substantial increases in threshold for diotic clicks but had little effect on detectability of monaural stimuli" (Berg, 1985, p.

404). Berg (1985) also cites reaction time data (Burbeck and Luce, 1982), data showing differences between detection and recognition tasks (Macmillan, 1971, 1973), and auditory neuron response properties in support of the sustained-transient hypothesis.

Van Wieringen and Plos (1995a, 1995b) have examined the roles of rate, duration, and frequency extent in the perception of /ba/, /da/, /ab/, and /ad/, and found that extent, rather than rate, served as the primary cue for consonant identification. Our model proposes a physiologically motivated transient channel in which both extent and rate have an influence. In addition, the dependence on rate and extent can also code transition duration as an emergent property in certain parameter ranges. Much work remains to be done to clarify the role of these covarying properties of formant transitions and their representation by the auditory system.

5. PHONET Model: Gain–Controlled Working Memories and Chunking Networks

Perceptual effects wherein a later-occurring vowel context influences the percept of the initial consonant, are explained using a short-term storage mechanism for the model sustained and transient cell activities. In particular, the cell responses form distributed patterns of activation that are stored briefly in working memories. These sustained and transient working memories set the stage for feature-based, phonemic classification or categorization. The total pattern of stored activity across the sustained and transient channels is assumed to be mapped into phonetic categories via a competitive learning or self-organizing feature map network (Cohen and Grossberg, 1986, 1987; Grossberg, 1976, 1978, 1982, 1986; Grossberg and Stone, 1986b; Kohonen, 1984; Rumelhart and Zipser, 1985). Such pattern learning tunes the synaptic weights between the lower-level working memory features and the higher-level category neurons in proportion to the relative activity levels in the working memories.

Categories hereby come to encode the *ratios* of feature node activations across working memory. Sensitivity to the ratios, rather than the absolute magnitudes, of featural activities has a geometric interpretation as sensitivity to the direction, rather than the magnitude, of the feature vector. That is, when the input is filtered through the bottom-up weights, the resulting signal is proportional to the cosine of the angle between a vector made up of input activities and a vector made up of the synaptic weight strengths. For a given synaptic weight vector, relative input activities determine the direction of the input vector and hence the angle of match. Grossberg (1978, 1986) proposed that this natural sensitivity to direction, or ratio information, could provide an invariance that preserves phonetic percepts over variable speaking rates. The model proposed here similarly assumes that phonetic percepts are associated with the ratio of sustained to transient channel activity, which will be called the *S/T ratio*. We propose that subjects' probabilistic identifications are dependent on a value directly proportional to this internal state variable. In the case of the Miller and Liberman (1979) and Schwab *et al.* (1981) experiments, two CV stimuli that in the model generate the same S/T ratios predict the same probability of a /ba/ percept.

The Miller and Liberman (1979) experiment shows that the subject can be given two stimuli with identical initial formant transitions and will identify the stimuli based on the duration of the following vowel. In the context of the model, we would observe identical transient channel responses but differing sustained channel responses, and identify on that

PHONET MODEL   **DATA PREDICTION**

(A)   **(B)**
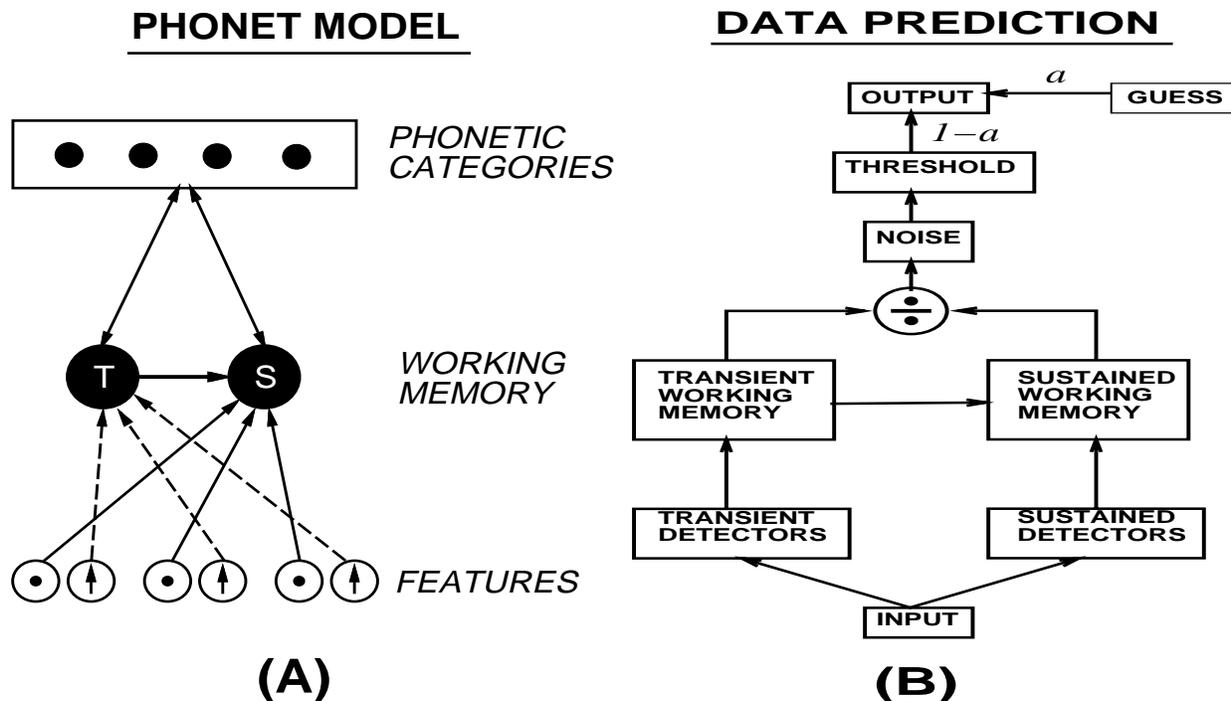
Figure 3: (A) PHONET Model. Formant transitions input acoustic energy across an array of local channels. At each spatial location, there is a pair of local feature detectors, one responsive to sustained input energy (circled dots) and the other to changes in energy (circled up-going arrows). The activations from the local detectors across a broad region are temporarily stored by sustained and transient channel working memories ("S" and "T"). The transient channel response modulates the processing rate of the sustained channel. (B) Data Prediction. Activities of S and T detectors are determined based on the input, and used to compute activities of the S and T working memories. The S/T ratio is formed from working memory activities in order to compute the same type of invariant as does the adaptive filter from working memory to phonetic categories, as in (A). White noise is added to the result. With probability $1 - a$ the subject responds /ba/ if the resulting sum exceeds a threshold, and with probability a, the subject guesses.

basis. On the other hand, two stimuli presented at different overall rates can have different vowel and/or formant transition characteristics and still generate the *same* phonetic percept due to model mechanisms which ensure that the S/T ratio remains invariant.

The Schwab *et al.* (1981) data shows that two stimuli with identical subsequent vowel durations and formant transition rates can produce different percepts if the formant transitions differ in extent (and therefore duration). On the assumption that the transient channel is sensitive only to the rate of the formant transitions, the model would produce identical outputs for two such stimuli. However, the transient channel model is proposed below as a network of transient working memory cells arranged in an on-center, off-surround, or lateral inhibitory, architecture, an organization characteristic of cortical tissue (Nabet and Pinker, 1991; Pickles, 1988; Shamma and Symmes, 1985). Inhibition from rate-sensitive cells in inhibitory surround frequencies allows the transient channel to code also the *extent* of frequency

transitions. The model thus assigns a key functional interpretation to competitive networks of onset detectors in coding speech transitions. Intuitively, as the sweep passes through successive frequency bands, it excites rate detectors centered at a number of different best frequencies. Each transient channel cell is inhibited in proportion to the activity of the sum of these cells centered at frequencies in its inhibitory surround. With a broad inhibitory surround, the net inhibition is proportional to the frequency extent of the sweep. The model below shows that this mechanism, namely a transient channel excited by faster rates but inhibited by broader frequency extents, produces a measure predictive of the perceptual data. The relevant type of inhibition occurs in neurons that obey membrane equations (Hodgkin, 1964). It is a *shunting* type of inhibition that tends to normalize activities across the network (Grossberg, 1973).

The perceptual invariance of the S/T ratio as speech rate changes suggests that an interaction occurs from the transient channel to the sustained channel. To see how the channels should interact, consider the responses to a CV stimulus as the syllable duration is shortened while the relative durations of the transitions (of fixed extent) and steady-state formants remain constant. Under the hypothesis of transient and sustained signal detectors, the transient and sustained channel responses behave inversely. The transient channel sees a faster formant transition and thus reacts more vigorously. The sustained channel sees a shorter vowel, has less time to integrate, and generates a smaller response. Yet the Miller and Liberman (1979) data provides examples of such stimuli yielding the same percept. The ratio of these responses thus changes even while the percept is predicted to be invariant. If, however, the transient response modulates the input to the sustained channel by changing its averaging rate, or gain, then the faster transitions would boost the incoming vowel signal, causing the sustained response to grow faster. In this way, cross-stream $T \to S$ gain control can create a more rate-invariant (T, S) working memory representation from rate-variant inputs to the T and S processing streams. A schematic diagram of the model architecture outlined above is given in Figure 3.

6. Modeling Methodology for PHONET

Cells in the transient and sustained working memories will herein be assumed to respond only to formant transitions and steady-state formants (vowels), respectively. These working memory cells store the activities of FM rate-sensitive cells and "buildup" cells, that integrate over the duration of a tone, in the cochlear nucleus (Britt and Starr, 1976a, 1976b; Rhode and Smith, 1986). The transient cell activities, denoted by T_f, are excited by transient detectors responsive to rates of stimuli passing through their (narrow) frequency range $f \pm \Delta f$ (Britt and Star, 1976b; Mendelson, *et al.*, 1993; Møller, 1983; Tian and Raushecker, 1994). The more rapidly a formant transition sweeps through the frequency band of a given transient cell, the greater the response of that cell. Because the transient cells are arranged in a competitive network, though, a given transient cell activity T_f is inhibited by a formant transition which excites other transient cells centered at frequencies $g \neq f$. Thus, all other things being equal, the response T_f will be less to a formant transition passing through a large range of frequencies (and thereby exciting many other transient cells centered at frequencies $g \neq f$) than to a transition at the same rate passing through a narrow range of frequencies centered at f.

Sustained cells S are excited by sustained detectors whose response is approximately constant over the duration of a steady-state tone (Britt and Starr, 1976a; Rhode and Smith, 1986). The sustained working memory then integrates the output of these detectors over time. The automatic gain control from the transient channel can alter this integration rate in a context-sensitive way.

7. PHONET Model Equations

Transient Working Memory

Transient working memory cell activities T_f obey membrane equations (Hodgkin and Huxley, 1952) embedded within an on-center off-surround network (Grossberg, 1973). As a result, excitatory and inhibitory inputs are multiplicatively shunted. Each cell in the transient working memory is excited by rate-sensitive transient detectors R_f that respond to linear FM sweeps in a narrow frequency band around each frequency f and by a source of tonic excitation R_{tonic} (Shamma and Symmes, 1985; Zurita *et al.*, 1994). The transient working memory cells have broad inhibitory sidebands, spanning a frequency range F assumed to encompass the maximal frequency extent of the formant transitions used as experimental stimuli. Cell activity T_f passively decays at rate a and is bounded by finite activities b and $-c$, respectively. This yields the following activity equation for the transient working memory activities:

$$\epsilon \frac{d}{dt} T_f = -aT_f + (b - T_f)(R_f + R_{tonic}) - (T_f + c)(\sum_{g \in F} f(R_g)) \tag{1}$$

In (1), the response of the transient detectors R_f is assumed to be proportional to the local rate of the input stimulus. Thus, if the stimulus is a linear frequency sweep with constant rate R over a set of frequencies $g \in F$, then the response R_f of the transient detector centered at frequency f is given by $R_f = k_r R$, where K_r is constant. For simplicity, each transient detector R_g is assumed to contribute a constant amount of inhibition $f(R_g) = k_e$ to the neurons in the transient channel. In particular, the signal function f generates a rapidly saturating response to R_g. Since each such detector is excited by energy in a narrow frequency band $g \pm \Delta g$, their sum over the broad frequency region F is proportional to the extent of the sweep that excites that transient channel. Assuming that the frequency range F contains the transition extent E, the total inhibitory signal in (1) satisfies

$$\sum_{g \in F} f(R_g) = k_e E. \tag{2}$$

To ensure that T_f reflects the instantaneous slope of the formant transitions, ϵ is chosen small in (1). Then T_f's response rate, ϵ^{-1}, is very fast. With these assumptions, when the transient channel is stimulated by a formant transition with constant rate R and frequency extent $f_{max} - f_{min} = E$, Eqn. (1) rapidly equilibrates to the value

$$T_f = \frac{b(k_r R + R_{tonic}) - c k_e E}{a + k_r R + k_e E}. \tag{3}$$

These activities are assumed to be stored in the transient working memory. Since, under

these conditions, the equilibrium transient activity T_f is same at all activated frequencies f, we may take $T = T_f$ as the output of the transient working memory.

Sustained Working Memory

The sustained channel working memory receives signals σ_f from sustained detectors tuned to narrow frequency bands $f \pm \Delta f$, assumed to obey the equation $\sigma_f = 1$ if a stimulus is present in frequency $f \pm \Delta f$ and $\sigma_f = 0$ otherwise. The sustained channel is also gain-controlled by the transient channel through a gain function, $g(T)$. Thus, the activation of cells in the sustained channel working memory, S_f, is given by

$$\frac{d}{dt}S_f = H(\sigma_f)g(T)[-S_f + \sigma_f], \tag{4}$$

where the Heaviside function H also gates integration on in the presence of sustained input and off it its absence. This provides a simple way to represent storage of sustained activities in a working memory. The term $-S_f$ represents the passive decay of sustained channel activity. Term $-S_f$ also allows S_f to track σ_f through time.

We quantitatively fit the Miller and Liberman (1979) under the simplifying assumptions that

$$g(T) = T = \frac{b(k_r R + R_{tonic}) - ck_e E}{a + k_r R + k_e E} \tag{5}$$

and that term $(a + k_r R + k_e E)^{-1}$ can be approximated by a Taylor expansion to first order. Since

$$(1 + x)^{-1} \approx 1 - x + O(x^2), \tag{6}$$

we have to first order

$$(a + k_r R + k_e E)^{-1} \approx 2 - a - k_r R - k_e E. \tag{7}$$

Thus, letting $R_T = R_{tonic}$,

$$g(T) \approx (b(k_r R + R_T) - ck_e E)(2 - a - k_r R - k_e E)$$

$$= (2 - a)bR_T + bk_r(2 - a - R_T)R - k_e(2c - ac + bR_T)E + k_e k_r(c - b)RE - bk_r^2 R^2 + ck_e^2 E^2. \tag{8}$$

Tests using this approximations showed that the quadratic terms in E^2 and R^2 made only a negligible contribution to the model fits (see Appendix C). Keeping only the linear and cross terms, this approximation gives the four parameter function of rate and extent

$$g(T) = g_1 + g_2 R + g_3 E + g_4 RE. \tag{9}$$

While best results were obtained with the cross-channel gain function given by Eqn. (9), we tested a variety of related expressions for g (see Appendix C). Of particular interest is the case where $g(T) = $ constant for all T. In this case, the gain-control of the sustained channel is independent of transient channel activity, so there is no cross-channel interaction *per se*. As shown below, PHONET performs significantly worse without the cross-channel gain-control interaction embodied by Eqn. (9).

Model Output

We can solve the two equations (3) and (4) explicitly for the S/T ratio, , $= S/T$, in terms of the stimulus variables R (transition rate), E (transition extent), and V (vowel duration), at the time the syllable ends. Note that T has equilibrated at the end of the transition interval, when the vowel begins, and that S only integrates its input over the vowel interval. Letting the sustained input, σ_f, be a constant equal to unity during the vowel (i.e., from $t = 0$ to $t = V$), the response of the sustained activity in (4) to a vowel with formant frequencies f_i is given by $S_{f_i} = 1 - e^{-g(T)V}$ and $S_g = 0, g \neq f_i$. Since the activities of all sustained working memory cells that are excited by a given vowel are equal, we take $S = S_{f_i}$ as the measure of sustained working memory activity.

At vowel offset, the response ratio is found by integration of (3) and (4) to be

$$, (R, E, V) = \frac{S}{T} = \frac{1 - e^{-g(T)V}}{\frac{b(k_r R + R_T) - c k_e E}{a + k_r R + k_e E}}$$

$$= \frac{a + k_r R + k_e E}{b(k_r R + R_T) - c k_e E} \left(1 - e^{-g(T)V} \right). \tag{10}$$

Thus , takes the form of a quotient of linear functions of transition rate and extent multiplied by an exponential function of vowel duration.

The ratio , is transformed into a probability P of responding /ba/ by passing it through a Gaussian cumulative distribution function Φ whose mean μ is a free parameter, and taking a linear function of the result:

$$P = \alpha_1 \Phi_\mu(,) + \alpha_2 \tag{11}$$

where

$$\Phi_\mu(x) = 1/\sqrt{2\pi} \int_{-\infty}^{x} e^{-(t-\mu)^2/2} dt. \tag{12}$$

The parameters α_1 and α_2 are subject to different interpretations in the two data sets. For the Miller and Liberman (1979) experiment, they can be interpreted as defining a mixture model. With probability α_1 the subject attends to the stimuli and uses , to determine the probability of response. On the remainder of the trials the subject guesses /ba/ with probability $\frac{\alpha_2}{1-\alpha_1}$. These parameters, suggestive of an attentional mechanism operating in the experiments, are necessitated by the data which asymptotes at probabilities greater than zero and less than one. For the Schwab *et al.* (1981) experiment, however, α_1 and α_2 are treated as a rescaling that converts rating to probability using Eqn. (11). Since the Schwab *et al.* (1981) data is a set of mean ratings $r \in [1, 2, \cdots, 6]$, it is assumed that the average rating on a particular trial can be predicted as a linear combination of the underlying probabilities. To fit the rating data, they were rescaled to be between zero and one by the formula $p^* = (r - 1)/5$. These numbers were then treated as if they were probabilities for the purposes of fitting the data.

To test whether the reliability of the predictions of these models was due to the relative simplicity of the data set, or to the detailed properties of the model, a simple *Null Hypothesis Model* was constructed. This model assumes that the transition duration, rate, frequency extent, and vowel duration are measured and a linear sum of these factors plus a constant bias term is determined. This output is added to a Gaussian random variable with unit

variance. If the sum is positive, the subject reports /wa/ and otherwise reports /ba/. This model takes the form

$$p = n_1\Phi(n_4R + n_5E + n_6V + n_7D + n_3) + n_2, \tag{13}$$

where as above R is the rate of the F2 formant transition, E is its extent, D its duration, V is the subsequent vowel duration, and n_i are parameters to be determined.

8. Model Simulations

This and subsequent subsections describe several statistical tests that probe how well the PHONET model fits the data and systematically study the deviations of the model's predicted values from the data. Figures 4 through 7 show the fits of the model to the Miller and Liberman (1979) and the Schwab *et al.* (1981) data. The qualitative fits to the data are good in all cases, although the combined optimization produces slightly worse fits on both data sets. PHONET was tested by directly fitting the ensemble of psychometric functions produced by Miller and Liberman (1979) and further fitting the data of Schwab *et al.* (1981) without parametric change. To construct a direct fit to the Miller and Liberman (1979) experiments, we convert the ratio of the sustained and transient channel activities to probabilities for the Miller and Liberman (1979) data. For the Schwab *et al.* (1981) data, we assume that a fixed linear transformation relates the average rating to the probability that would be generated for the Miller and Liberman (1979) data.

The probabilities are generated from the S/T ratios in the following manner. A fixed proportion of the time the subject attends to the input data. During these trials, the ratio of the sustained and transient channel activities , is added to a Gaussian random variable $\mathcal{N}$ of fixed mean and unit variance. A positive result yields a response of /ba/ while a negative result yields a response of /wa/. In the remainder of the trials the subject guesses, providing either response on approximately 50% of the trials. PHONET's performance is also compared to the Null Hypothesis model and to several variants of PHONET that assess the predictive contribution of different combinations of acoustic parameters. Section 8.1 discusses the fit of the model to the data and shows that PHONET captures most of the variation in the data. In addition, it is shown that for the Miller and Liberman (1979) data, the model predicts and the data confirm that there is a ceiling on the likelihood of hearing /ba/ as well as a floor on the likelihood reported of hearing /wa/. This is consistent with an attentional mechanism operating in these experiments.

To decide whether the model's outstanding goodness of fit results are due to the simplicity of the data set, the Null Hypothesis model is compared with the data in Section 8.2. This model fits the Schwab *et al.* (1979) data about as well as the original model. However, the fit to the Miller and Liberman (1979) data is far worse than PHONET's fit. The simplified model fails to capture the change in the shift of the threshold with vowel duration that the PHONET model captures easily. Section 8.3 then examines variants of PHONET which retain PHONET's exponential dependence on vowel duration, but use alternative transient channel activity equations. Finally, Section 8.4 probes the small discrepancies from PHONET's fits to the data by a systematic study of the model residuals. In particular, the residuals are plotted against the model prediction to see if there is any bias or trend in the deviations of the model from the data.
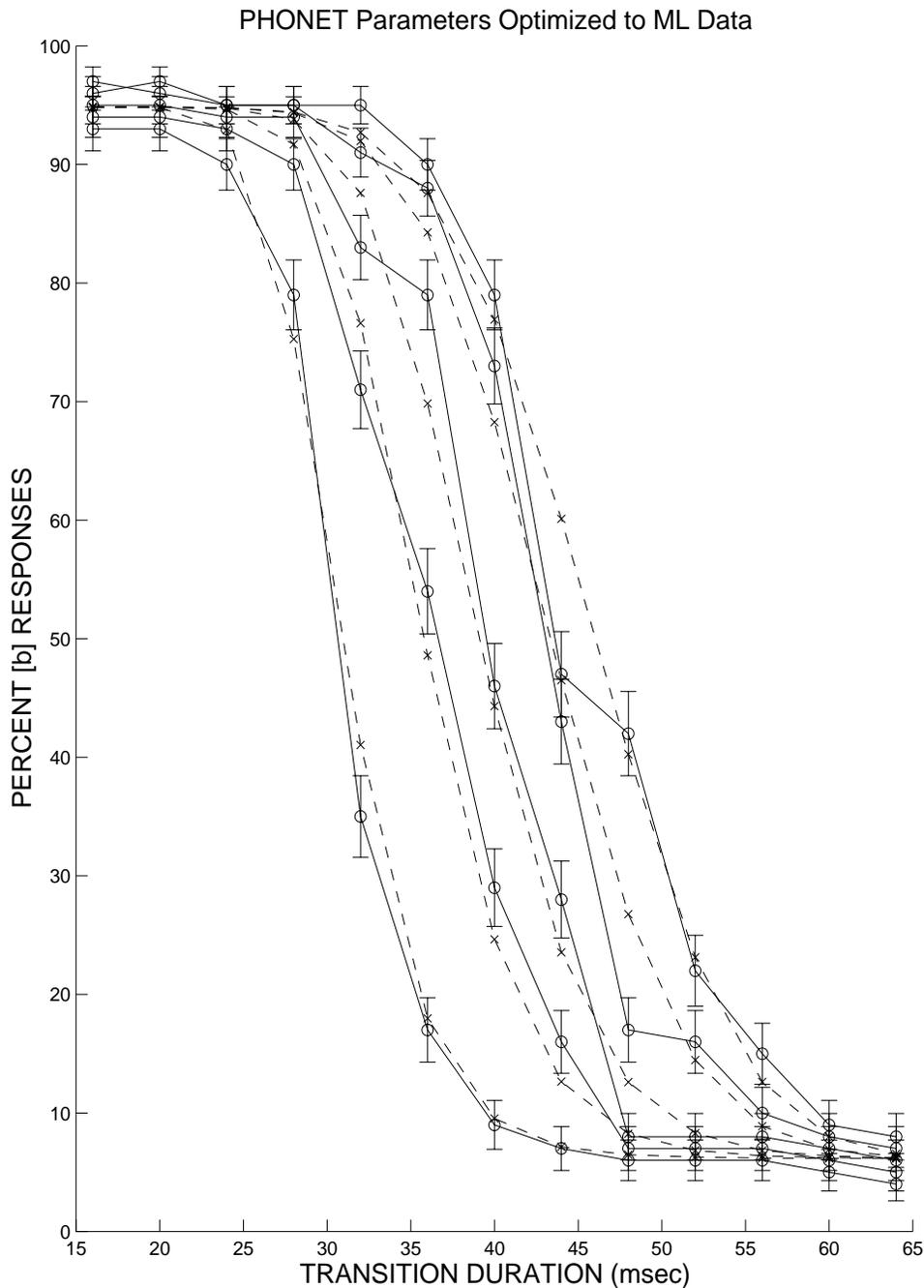
Figure 4: Fit of the S/T model to the Miller and Liberman (1979) data. The data are shown in solid lines, and the model probabilities are shown in dashed lines. Model parameters were optimized to fit the Miller and Liberman data only, and are listed in the middle column of Table 1.
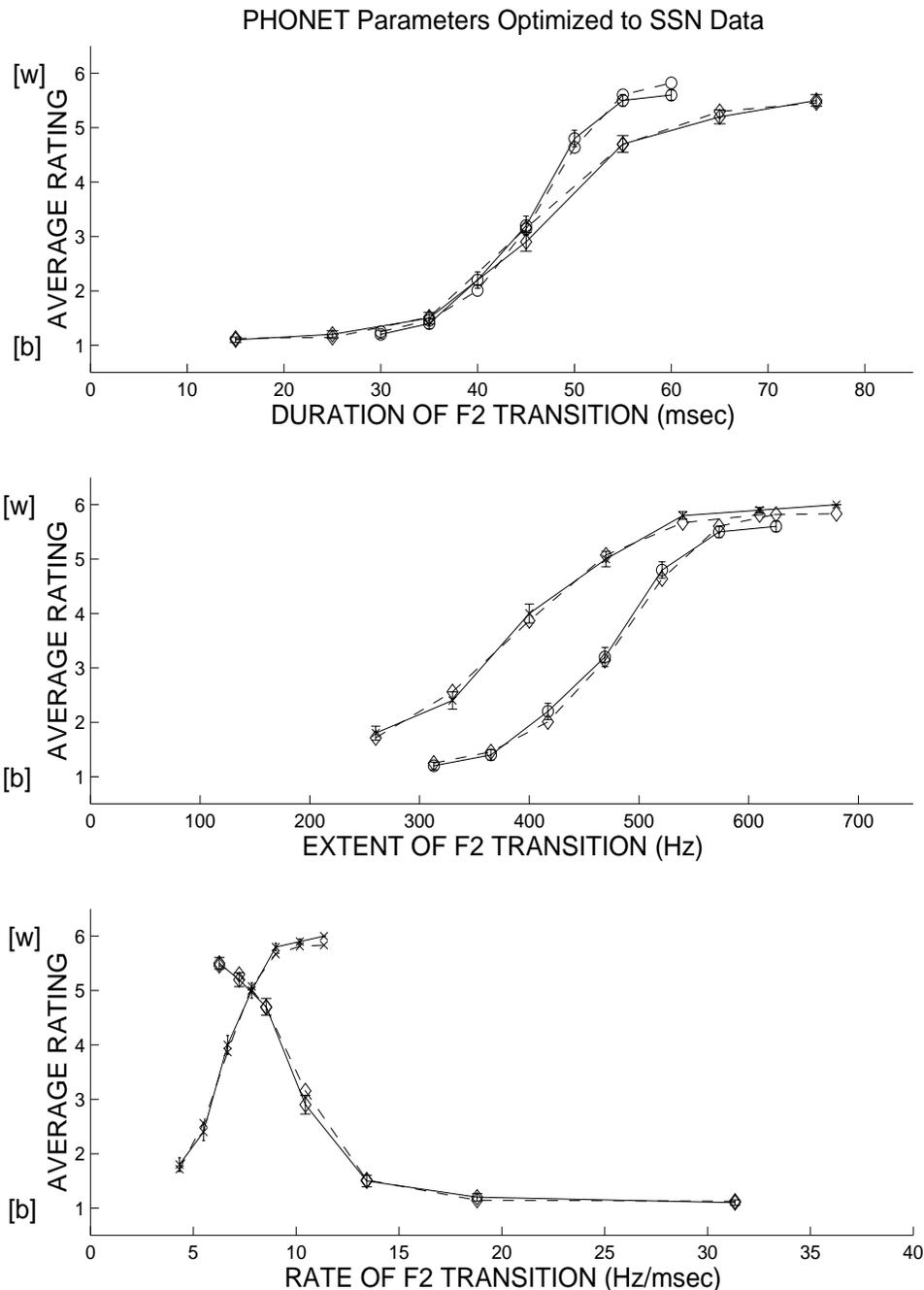
Figure 5: Fit of the S/T model to the Schwab, Sawusch, and Nusbaum (1979) data. The data are shown in solid lines, and the model probabilities are shown in dashed lines. Model parameters were optimized to fit the Schwab *et al.* data only, and are listed in the left column of Table 1.

Name	SSN Fit	ML Fit	Combined Fit
α_1	0.9411	0.8871	0.8899
α_2	0.0324	0.0611	0.0583
μ	0.3438	2.411	1.59
a	3.01	1.07×10^{-6}	4.60
k_r	0.0048	0.0360	0.0135
k_e	2.4×10^{-14}	1.2×10^{-14}	9.4×10^{-12}
ck_e	1.04×10^{-6}	4.7×10^{-7}	1.5×10^{-7}
bk_r	4.55×10^{-12}	0.0026	0.006
bT_{tonic}	1.001	0.8854	1.461
g_1	0.0024	0.0055	0.0076
g_2	4.15×10^{-5}	7.714×10^{-5}	3.97×10^{-5}
g_3	3.85×10^{-5}	-6.90×10^{-5}	-6.71×10^{-5}
g_4	1.48×10^{-6}	6.01×10^{-7}	3.56×10^{-7}

Table 1: Parameters for the Several Optimizations. SSN=Schwab, Sawusch, and Nusbaum (1981); ML=Miller and Liberman (1979).

8.1 Parameters and Fit of the Model to the Data

The model predictions are probabilities p_i for the Miller and Liberman (1979) data and a set of rescaled ratings r_i for the Schwab *et al.* (1981) data. The r_i are treated as if they were true probabilities predicted via the models. Parameters are optimized separately for each experiment and the two experiments combined using maximum likelihood estimation (Rao, 1973). Using Stirling's Formula to approximate $N!$ for a binomial distribution, maximizing the log likelihood, and therefore the likelihood, of the reported data is equivalent to minimizing

$$\Lambda_\Theta = 2N_i \sum_i p_i^* \log\left(\frac{p_i^*}{p_i}\right) + q_i^* \log\left(\frac{q_i^*}{q_i}\right), \tag{14}$$

where i indexes the stimulus, N_i is the number of recorded responses to each stimulus, p_i is the predicted probability, p_i^* the reported probability, $q_i = 1 - p_i$, and Θ is the vector of model parameters. Λ_Θ is known to be asymptotically distributed as a χ^2 random variable with $S - P$ degrees of freedom, where S is the total number of data points and P is the number of parameters (Rao, 1973). The parameters were optimized by the Nelder simplex method (Press *et al.*, 1988). Table 1 lists the parameters for the separate and combined optimization.

The parameters $a, b, T_{tonic}, c, k_e, k_r, \mu, \alpha_1, \alpha_2, g_1, g_2, g_3$ and g_4 are defined in Eqns. (9) through (11). These parameters are roughly similar for each of the three conditions of optimization. However note that parameters α_1 and α_2 define the low and high asymptote of the probabilities via equation (7): α_1 is near one for the Schwab *et al.* (1981) data, but
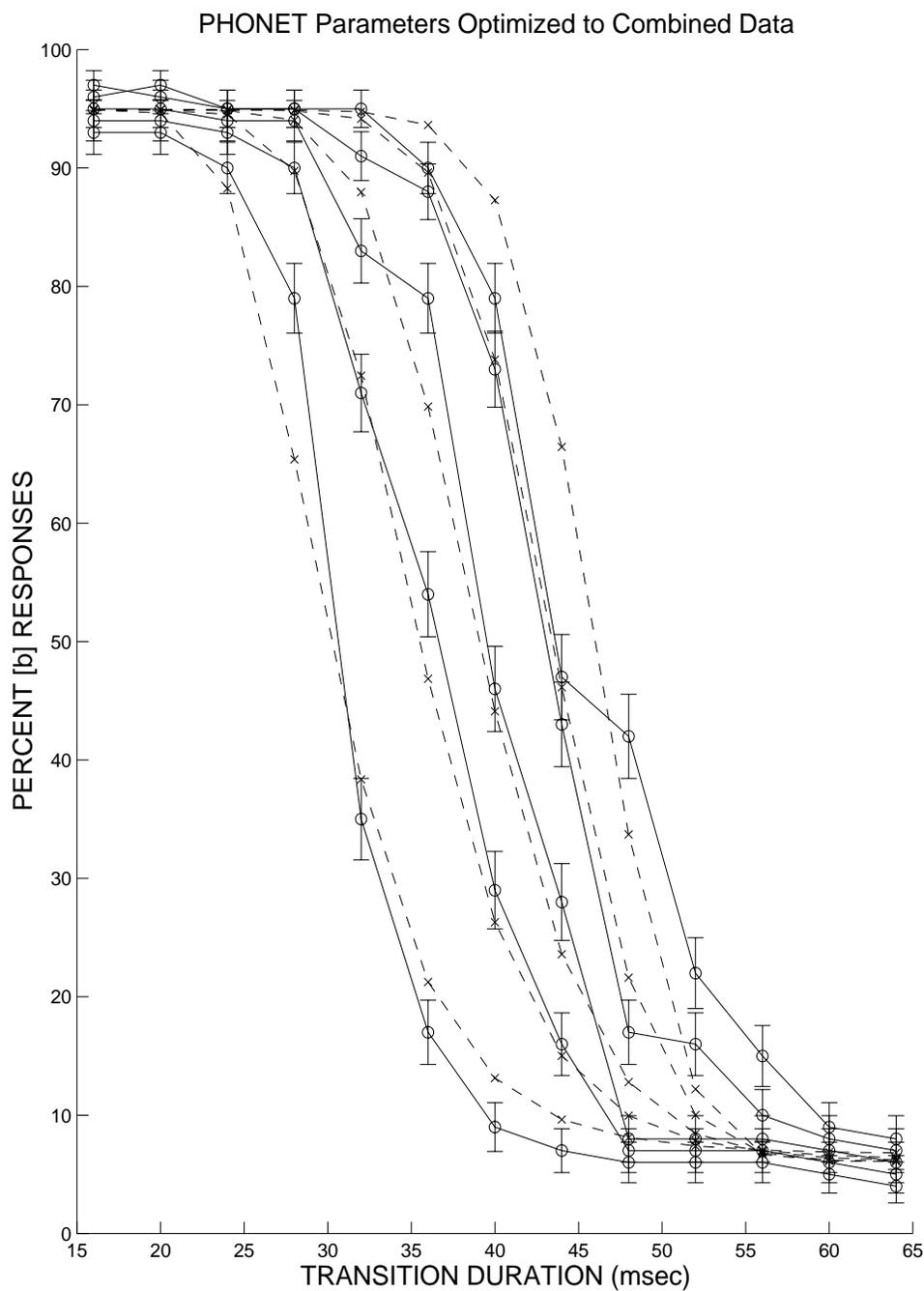
Figure 6: Fit of the S/T model to the Miller and Liberman (1979) data, using parameters optimized for a simultaneous fit of both data sets. Data shown in solid lines, model probabilities in dashed lines. Model parameters are listed in the right column of Table 1.

is about 0.89 for the Miller and Liberman (1979) data. This discrepancy exists because of the multiple use of these parameters. In the case of Schwab *et al.* (1981), these parameters define the rescaling from predicted probabilities to predicted ratings. In the case of Miller and Liberman (1979), they define the asymptote of predicted probabilities. The parameters α_1 and α_2 are consistent with the attentional hypothesis described above for the Miller and Liberman (1979) data. Under this hypothesis, the subjects of these experiments attend to the stimulus only approximately 89% of the time. The other 11% of the time they simply guess with a bias favoring /ba/. If this is the case, a serious attempt to decide when the subjects are attending to the stimulus or to reward performance when they attend may decrease the high and low asymptotes of these curves.

Table 2 shows goodness of fit statistics associated with the model. The rows of the table refer to the statistics measured from the model output and data. The first column refers to parameters optimized for the Schwab *et al.* (1981) data only, the second column to parameters optimized for the Miller and Liberman (1979) data only, and the third shows the results of the combined optimization. The probabilities reported are significance levels of the associated statistics.

Statistic	SSN Fit	ML Fit	Combined Fit
ρ^2	0.9950725	0.992907	0.988409
$F(n_1, n_2), p$	$119(13, 7), p < 10^{-6}$	$548(13, 51), p < 10^{-20}$	$472(13, 72), p < 10^{-20}$
$\Lambda, (N, p)$	$30, (N = 8, p = 0.0002)$	$84, (N = 52, p = 0.0032)$	$215, (N = 73, p < 10^{-16})$
χ^2	31	82	213
$\chi'^2, (N', p)$	*	59, (N=50, $p = 0.171$)	$78, (N = 60, p = 0.0539)$

Table 2: Goodness of Fit Statistics of PHONET. ρ^2 = proportion of variance predicted by model. $F = n_2 \text{var}(\text{model})/n_1 \text{var}(\text{residuals})$. $n_1 = P - 1, n_2 = S - P - 1, N = \#$ of degrees of freedom $= S - P$, where $S = \#$ data points, $P = \#$ parameters. $N' = S' - P$, where S' = reduced data set. Λ = defined in Eqn. 14. χ^2 = defined in Eqn. 15. * = Too few data points relative to parameters evaluate this statistic.

The first line shows the square of the total correlation coefficient of the model prediction with the reported data. This represents the proportion of the variance predicted by the model. Note the extremely high value. In the worst case PHONET predicts over 98.8% of the variance of the data.

F statistics were computed to test the hypothesis that the model predicts all this variance by chance. The second row of Table 2 shows the statistics, the degrees of freedom in the F test and the significance level of the test. In all three cases, the hypothesis that the model prediction is no better than chance is rejected at a significance level of less than 10^{-20}. While the statistical hypotheses underlying the F test assume normality, linear regression, and equal variance for each of the values estimated, the statistic is somewhat robust against violation of these hypotheses (see Scheffe, 1959). The very high level of significance and the highly accurate fit obtained argue strongly that the model accurately predicts the data, and the data are not predicted by chance.

PHONET Parameters Optimized to Combined Data
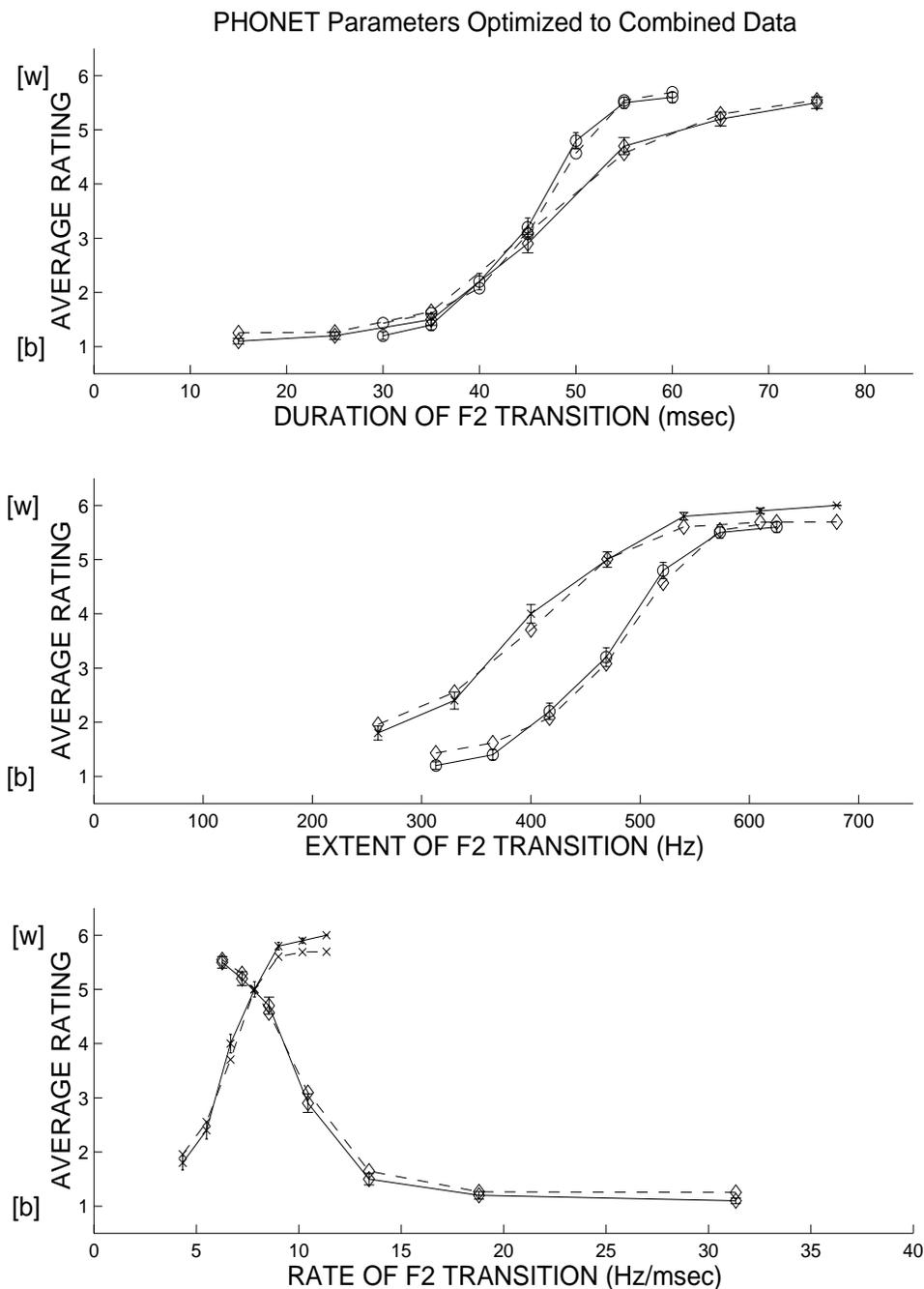
Figure 7: Fit of the S/T model to the Schwab *et al.* (1979) data, using parameters optimized for a simultaneous fit of both data sets. Data shown in solid lines, model probabilities in dashed lines. Model parameters are listed in the right column of Table 1. In comparison with Figure 5, the model gives a poorer fit than when parameters are chosen for the Schwab *et al.* data alone, but the model still predicts over 98% of the variance in the data.

The third row of Table 2 reports likelihood ratio statistics for the model, testing the null hypothesis that the deviances of the model from the reported data derive from the statistical nature of the model; i.e., that the predictions are within the noise level of the binomial distribution. The likelihood ratio is asymptotically distributed as χ^2 with the $S - P$ degrees of freedom, where S is the number of data points predicted and P is the number of parameters. As an alternative check on these significance levels, we report χ^2 statistics of the form

$$\chi^2 = \sum_i N_i (p_i - p_i^*)^2 / p_i q_i, \tag{15}$$

where p_i are the model predicted probabilities and p_i^* are reported probabilities or rescaled ratings, and $q_i = 1 - p_i$. The asymptotic theory of large sample statistics predicts that the statistics Λ_Θ and χ^2 would have the same asymptotic distribution if the maximum likelihood estimator of Θ constructed in equation (15) is used to obtain the values p_i in (14). The fact that the statistics reported in rows three and four of Table 2 agree to within statistical error suggests that the parametric estimate is close to the maximum likelihood estimate and that large sample theory and tests are applicable to these two experiments.

Approximating the χ^2 distribution on N degrees of freedom by a normal distribution with mean N and variance $2N$, it is apparent that the statistics given in row four of Table 2 are within a factor of two or three of accepting the hypothesis that the model predictions are within the noise level of the data. One possible reason for rejection is that the model does not well fit data points that are outliers to the trends reported in the data. These reported outliers may exist for many reasons, including the fact that the reported data were averaged across subjects.

To test this hypothesis, we chose the reported data points for which the model provided the poorest fit and expunged these from the database. Maximum likelihood estimators for the parameters were then constructed using the expunged database. For the Miller and Liberman (1979) data, only two data points needed to be excised in order to accept the χ^2 goodness of fit test at the $p = 0.05$ level. The much smaller number of points (n=21) in the Schwab *et al.* (1981) data test require a prohibitively small χ^2 statistic on 6 degrees of freedom in order to accept the test. Likewise, given the differing experimental conditions of that produced the two data sets and the different interpretations of the parameters, it is not surprising that the χ^2 test cannot be accepted for the combined data sets unless 13 data points are expunged. Still, this test reinforce the conclusion that the PHONET fits the Miller and Liberman (1979) data with an unusual degree of accuracy.

8.2 The Fit of the Null Model

One possible reason for the favorable goodness of fit statistics reported in Table 2 is the underlying simplicity of the data set. It is possible that many different models could adequately fit the data of Schwab *et al.* (1981) and Miller and Liberman (1979). The data of these experiments would then lack power to choose between models. To test this, the Null Hypothesis Model of Eqn. (12) was constructed to fit the data.

Table 3 shows goodness of fit statistics for the Null model. Observe that in both the cases of the combined data set and the Miller and Liberman (1979) data set, the proportions of variance predicted, ρ^2, are significantly lower than in the PHONET model. The likelihood

Statistic	SSN Fit	ML Fit	Combined Fit
ρ^2	0.992345	0.970182	0.966151
$\log \Lambda(N)$	41(13)	284(57)	452(78)
Z, p	0.65, $p < 0.26$	4.174, $p < 1.5 \times 10^{-5}$	2.96, $p < .0015$

Table 3: Goodness of fit statistics of the Null Model.

statistics are significantly larger as well, showing a poorer fit. The Null model fits the Schwab data, however, nearly as well as PHONET.

The ρ to Z transformation (Rao, 1973) was used to test whether one model produced a reliably better fit than the other model. It was assumed that the model with the higher correlation coefficient produces the true value of the statistical correlation of the data with the model, and asked how likely the better model would be in another experiment to produce a correlation coefficient smaller than the one calculated for the poorer model. If this occurs less than 5% of the time, then we can conclude the model with the higher correlation coefficient is a significantly better model.

Under these hypotheses, the statistic

$$Z = \sqrt{\frac{N-3}{2}}(\tanh^{-1}(\rho_1) - \tanh^{-1}(\rho_2)), \tag{16}$$

is Gaussian distributed with mean zero and variance one, where N = the number of degrees of freedom. The third row of Table 3 lists this statistic. For the Schwab *et al.* (1979) data, the fit produced by PHONET is better than that of the Null model but the difference is not significant. However, for both the Miller and Liberman (1979) data and the combined data set, the fit of the PHONET model is better and the difference is highly significant. The PHONET Model is superior to the null model and can be used to explain the data in all three cases.

Figures 8 and 9 show the fits to the data produced by the Null model. As predicted by the goodness of fit statistics, the fit to the Miller and Liberman (1979) data is much poorer than the fit by the PHONET model. In particular, the optimal fit fails to accurately capture the spacing of the different psychometric functions which the PHONET model accounts for easily. One reason for this is PHONET's use of the sustained channel to capture the exponential dependence on vowel duration.

8.3 PHONET Variants

To further elucidate PHONET's ability to accurately predict both data sets, several simple variants of PHONET were formed. These variants retained PHONET's sustained channel activity equation, shown via comparison with the linear model in Section 8.2 to importantly contribute to PHONET's fit. PHONET's expression for transient channel activity and the relatively small values of several transient channel coefficients, suggest alternative models for testing. By modifying the transient channel terms in Eqn. (10), various linear, multi-
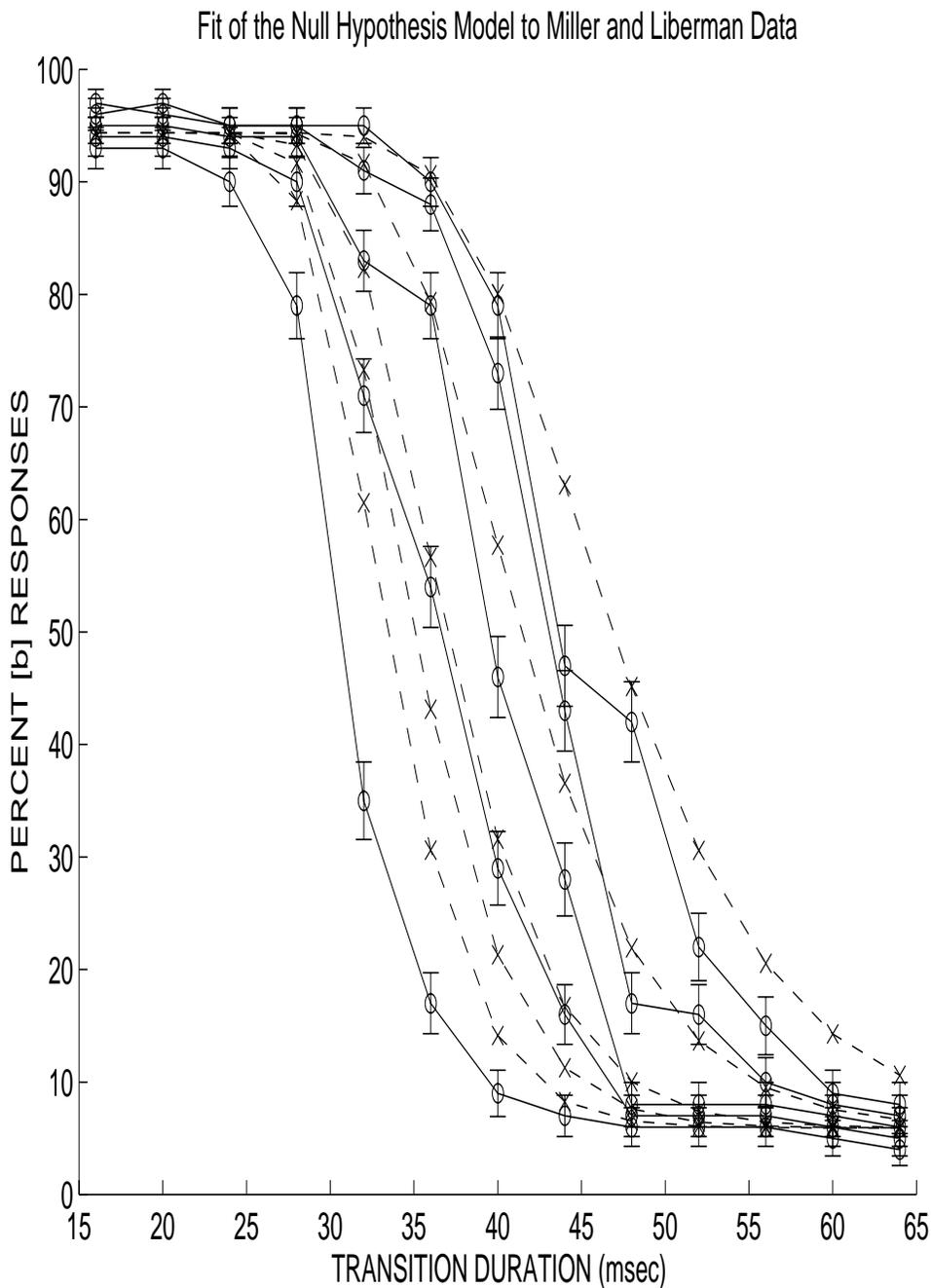
Figure 8: Fit of the Null Hypothesis model to the Miller and Liberman (1979) data. Data shown in solid lines, model probabilities in dashed lines. Note the inability of the model to accurately capture the dependence of response threshold on vowel duration.
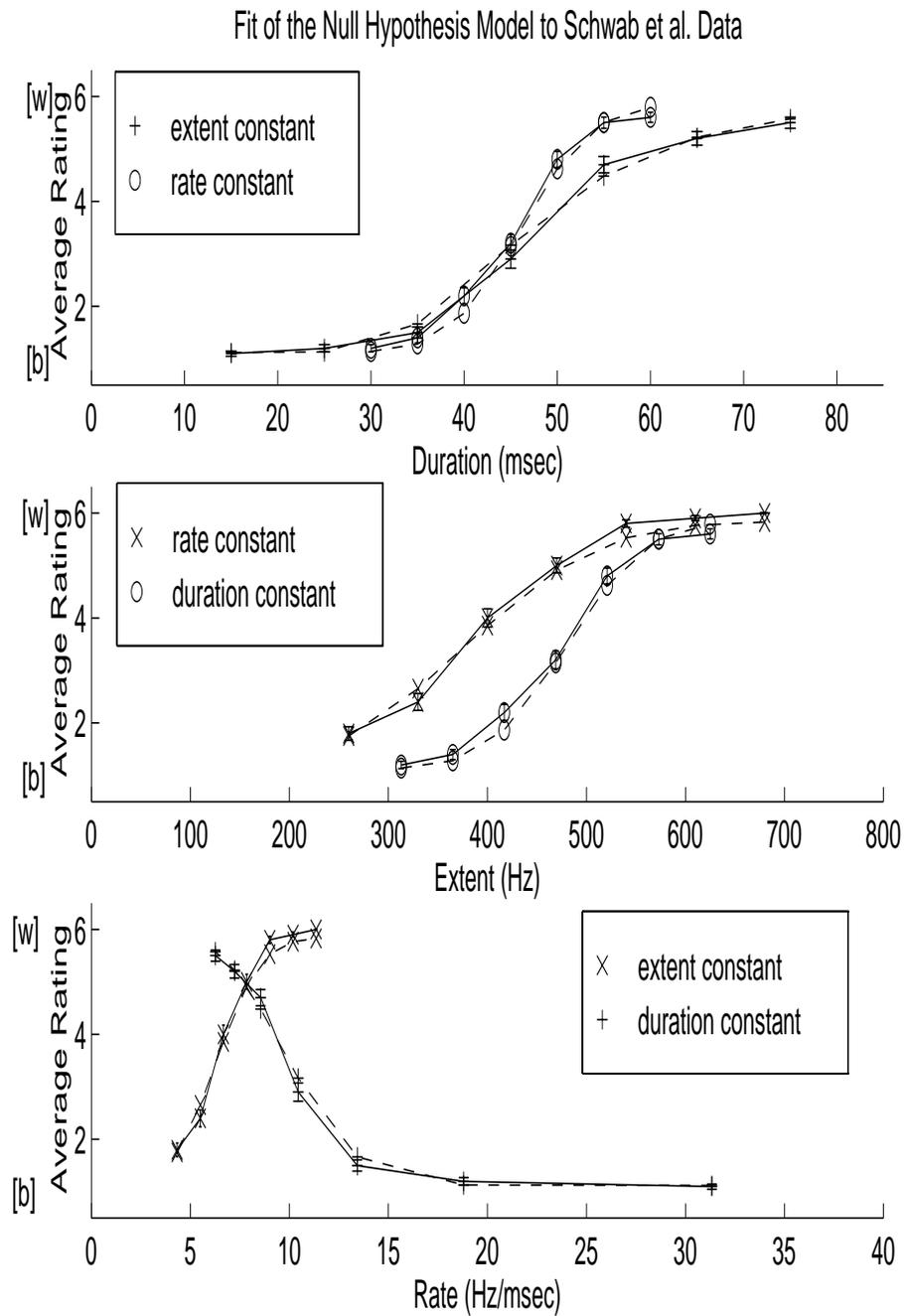
Figure 9: Fit of the Null Hypothesis model to the Schwab *et al.* (1981) data. Data shown in solid lines, model probabilities in dashed lines. The Null Hypothesis model fits this data set about as well as the S/T model.

plicative, and divisive combinations of transition duration, rate, and extent were tested. Of particular interest are a simple rate and extent combination (which retains PHONET's successful fits), and an extent and duration product (which fails to account for both data sets). While the variants explored in this section diverge from the neurophysiological foundation developed for PHONET, they offer some mathematical insight into the acoustic properties predictive of the data.

The term multiplying the output of the sustained channel in Eqn. (10) is equal to the reciprocal of transient channel output. The functional form of this term suggests that it may be possible to form a variant of PHONET by approximating the quotient of linear functions of rate and extent by a simple linear function, or a simple quotient. In particular, replacing Eqn. (10) by

$$, (R, E, V) = (\theta_1 + \theta_2 R + \theta_3 E) \left(1 - e^{-g(T)V}\right) \tag{17}$$

yields a simple approximation that performs significantly better than the form of PHONET in Eqn. (10) ($\rho^2 = 0.9925$ vs. $\rho^2 = 0.9884$, $p < 0.08$) when using the cross-channel gain function given by Eqn. (9). Moreover, with $g(T) = $ constant, the form given by Eqn. (17) still performs almost as well as PHONET using $g(T)$ given by Eqn. (9) ($\rho^2 = 0.9869$ vs. $\rho^2 = 0.9884$, $p = 0.65$). Likewise, modifying , to include the quotient of rate and extent given by

$$, (R, E, V) = \frac{R + \theta_1}{\theta_2(E + \theta_3)} \left(1 - e^{-g(T)V}\right) \tag{18}$$

also yields a functional form that performs almost as well as Eqn. (13) ($\rho^2 = 0.9854$).

Equation (18) may be compared to a transient channel equilibrium activity

$$T = \frac{b(E + R_T)}{a + R}. \tag{19}$$

This expression has three free parameters: passive decay rate a, tonic excitation R_T, and activity upper bound b. Eqn. (19) represents a transient channel *excited* by extent and *inhibited* by rate, an interpretation whose biological plausibility is not immediately evident. This transient channel representation of initial formant transitions as a function of frequency extent over rate is interesting in light of the relation rate = extent/duration. It is apparent that for certain parameter values,

$$\frac{b(E + R_T)}{a + R} \approx \frac{bE}{R} = bD; \tag{20}$$

i.e., the transient activity reflects a tuning to duration of the formant transition. Casseday, Erlich, and Covey (1994) have shown tuning to FM stimulus duration in bat inferior colliculus and have further established that the inhibition is mediated by the inhibitory neurotransmitter GABA. Casseday *et al.* (1994) did not dissociate the influence of FM sweep rate, duration, and extent in their experiments, but their physiological demonstration of complementary transient excitatory and inhibitory processes interacting to extract stimulus duration bears some similarity to the transient channel given in Eqn. (19).

Schwab *et al.* (1981) proposed that the product of frequency extent E of the formant transitions and their duration D be used as a variable to determine the likelihood of hearing

/ba/ or /wa/. While this may be a plausible heuristic, no fit to the data was provided. In this paper, two independent models not using this heuristic were proposed and fit to their data, explaining in both cases a proportion of variance in excess of 98%. Their data set, taken by itself, is thus not extensive enough to warrant their conclusion. We tested a PHONET variant which used the product ED in the transient channel term of , :

$$, (E, D, V) = (\theta_1 ED + \theta_2) \left(1 - e^{-g(T)V}\right) \tag{21}$$

While this model performed well on the Schwab *et al.* data in isolation, it failed to predict more than 97.78% of the combined data, significantly worse than PHONET ($p < 0.02$). Similar variants tested with different combinations of rate, extent, and duration revealed that rate and extent jointly perform better in all models than any other pairwise combination when a constant cross-channel gain function is used. When the four-parameter gain function of Eqn. (9) is used, many models produce fits that are statistically indistinguishable. Note that the gain function of Eqn. (9) includes rate and extent as the independent variables, and thus significantly improves the fit of models whose transient channel output does not include both rate and extent. For this reason, the constant gain function provides a finer gauge of the predictive contribution of the formant transition parameters to the transient channel models. Appendix C describes the other variants tested and presents the resulting statistics.

8.4 Study of Residuals: Model Discrepancy from the Data

In order to further compare the model fits to the data, the residuals between the model predictions and the data are plotted in Figures 10 and 11. Consider the Schwab *et al.* (1979) residuals, separately optimized and plotted in Figure 10. If it is the case that parametric estimation is unbiased there should be as many positive residuals as negative residuals. This is indeed the case for the Schwab data, which has 11 out of 21 positive residuals (Z = 0.44, $p < 0.67$).

If the estimates of the true value are unbiased, then the residuals should be asymptotically independent of the model predictions. Figure 10 shows that the slope of the regression line between the residuals and the model prediction is small. The slope of the regression line is not significantly different from zero ($\rho = -0.0003, Z = 0.001, p = 0.50$), and its mean is indistinguishable from zero.

It is desirable that the residuals have zero mean and show no trend in the placement of positive or negative errors. That is, the model prediction should be equally likely to overshoot or undershoot at any point with no sequence, or *run*, of overshoots or undershoots. The hypothesis that the residuals are unbiased rejects when there are too few runs, leading to residuals that overshoot when in certain data ranges and which undershoot systematically in others. If this is the case, then the model predictions show systematic bias when different outputs are predicted.

To test this, the number of runs n of positive (or negative) residuals was compared with the expected value and divided by the standard deviation. If the number of data points is relatively large, this statistic is asymptotically normal. The standardized deviate in this case ($Z = (n - E(n))/\sigma(n) = 1.13, p < .87$) fails to attain significance for the Schwab *et al.*
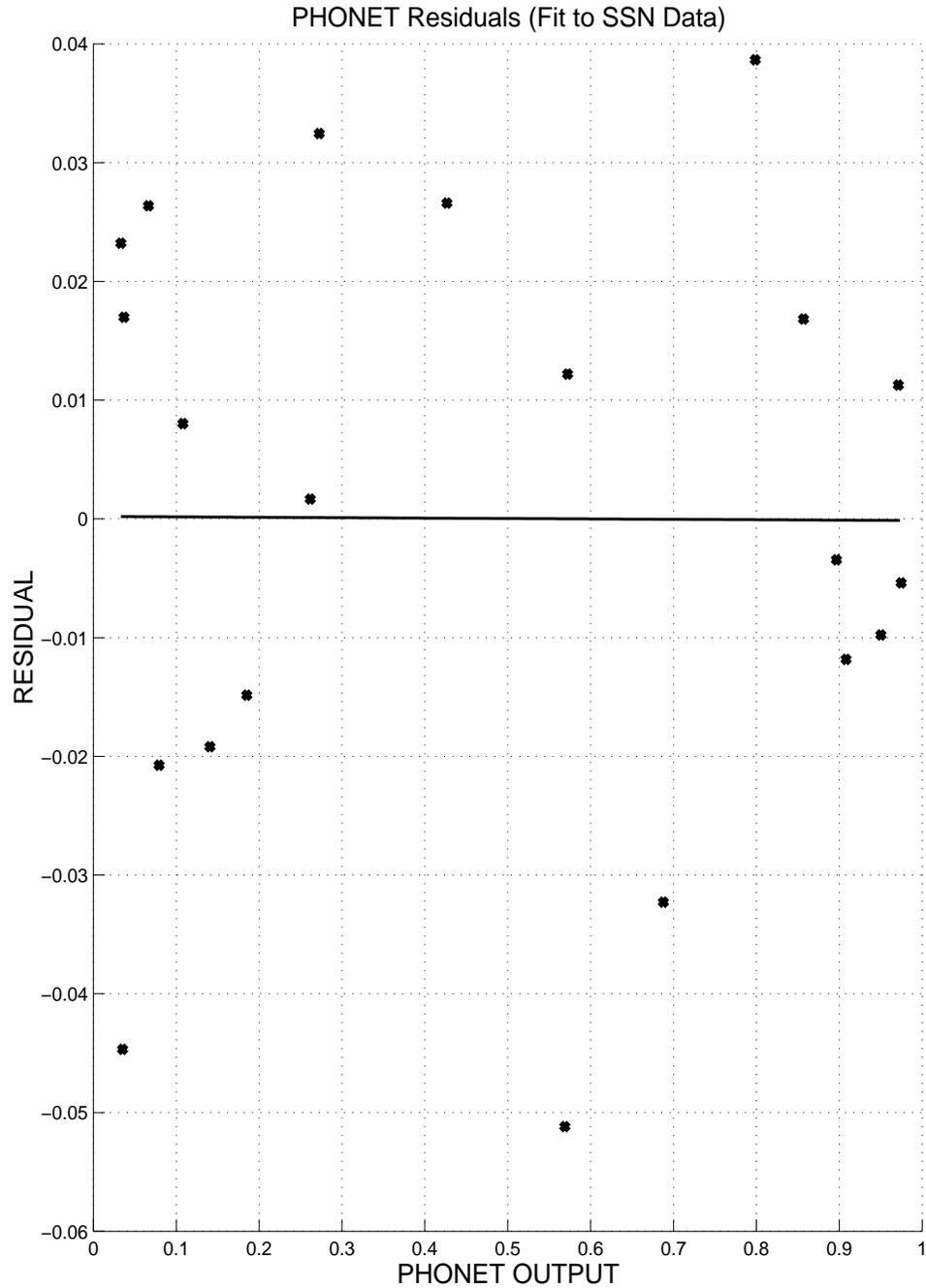
Figure 10: Residuals of the S/T model when fit to the Schwab *et al.* (1981) data plotted against model output. Parameters were optimized for the Schwab et al. data only. Residuals were computed by converting ratings to probabilities and subtracting from model output. The solid line shows the linear regression of residuals against model output. The residuals show no systematic trend and have the expected number of runs of positive and negative values.

residuals.

Now consider the Miller and Liberman (1979) residuals, which are plotted against the model predictions in Figure 11. Once again, neither the slope of the regression line ($\rho = -0.0011, Z = 0.006, p = 0.49$), the number of positive residuals ($Z = -0.49, p = 0.31$), nor the number of runs ($Z = 0.17, p = 0.57$) is greater than expected. Note that the residuals tend to be larger at the intermediate values than near zero or near one. This is, in part, a consequence of the parameter estimation procedure. The χ^2 statistic weights errors from the predicted values by the reciprocal of their variance, which grows at the tails of a binomial distribution.

Finally, when a single set of parameters is chosen for both data simultaneously, and the residuals of the entire data ensemble are plotted against PHONET's output, the model shows no systematic over- or underestimation of the probabilities as compared with the data. The hypothesis that the slope of the linear regression is zero ($\rho = -3.7 \times 10^{-5}, Z = -0.0002, p = 0.50$), the runs test ($Z = 0.67, p = 0.75$), and the sign test ($Z = -0.32, p = 0.37$) all accept, indicating independent, non-parametric confirmation that PHONET does not show systematic bias in its predictions.

9. Discussion: How the Brain Uses Ratio Information In Resonant Dynamics

In summary, the PHONET model produces a fit that approaches the the noise level predicted by the statistical model and explains over 98% of the variance in the data. The Null Hypothesis Model discussed above is rejected because it fits none of the data sets significantly better than the PHONET model, and fits the Miller and Liberman (1979) data set significantly worse. The ED Model, based on the criterion proposed by Schwab *et al.* also fails to fit the combined data. Other models inspired by PHONET that use different combinations of transition rate, extent, and duration in the transient channel are capable of fitting both data sets when the cross-channel gain function of Eqn. (7) is used. When a constant gain function is used many models can fit either the Miller and Liberman (1979) data or the Schwab *et al.* (1981) data, but the only models capable of fitting both data sets simultaneously are models whose output increases with rate and decreases with extent. Thus, a global result of this study is that careful statistical analysis can give a principled methodology for choosing between theories and showing where and how model predictions are discrepant from the data. In particular, choice between models, relative importance of rate, extent, and duration, the exponential dependence of model output on vowel duration, and the attentional mechanisms discussed above have resulted from this analysis.

The PHONET model shows how generic neural elements responsive to sustained and transient stimulus features can be combined to identify segments phonetically, with sensitivity to intrinsic temporal information. The model working memories operate at a pre-categorical stage, where auditory features such as transients and formants are represented, and its S/T ratio varies continuously with the sustained duration, and transient rate and frequency extent of input segments. Categorization based on analog working memory values is consistent with data on within-category identifications of speech tokens (Miller and Volaitis, 1989; Volaitis and Miller, 1992). For example, subjects' ratings of the *quality* of /pi/ tokens as exemplars of the stop consonant exhibit smooth, unimodal distributions with VOT. Both the
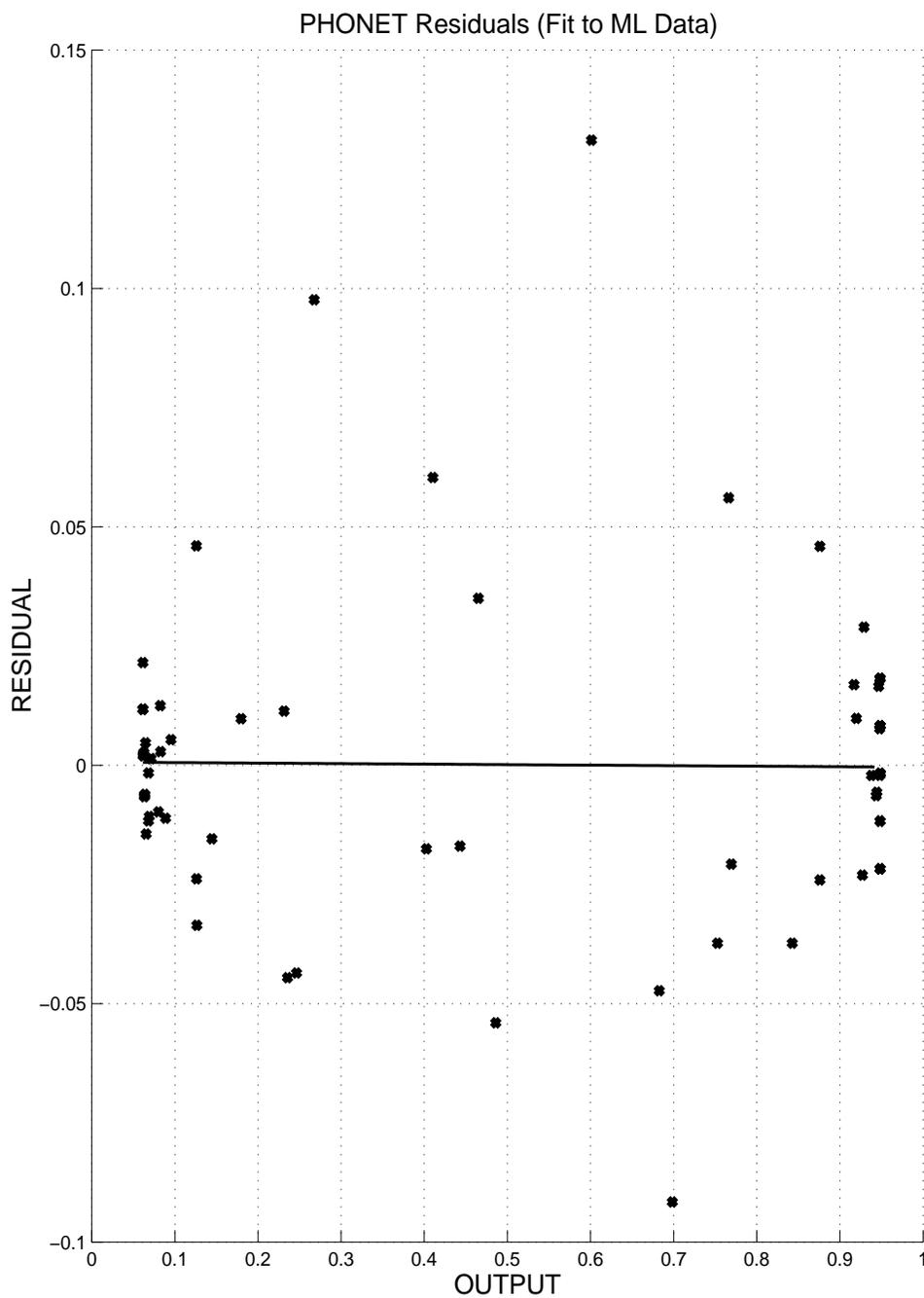
Figure 11: Residuals of the S/T model when fit to the Miller and Liberman (1979) data plotted against model output. Parameters were optimized for the Miller and Liberman data only. The residuals show no systematic trend, and show the expected signs.
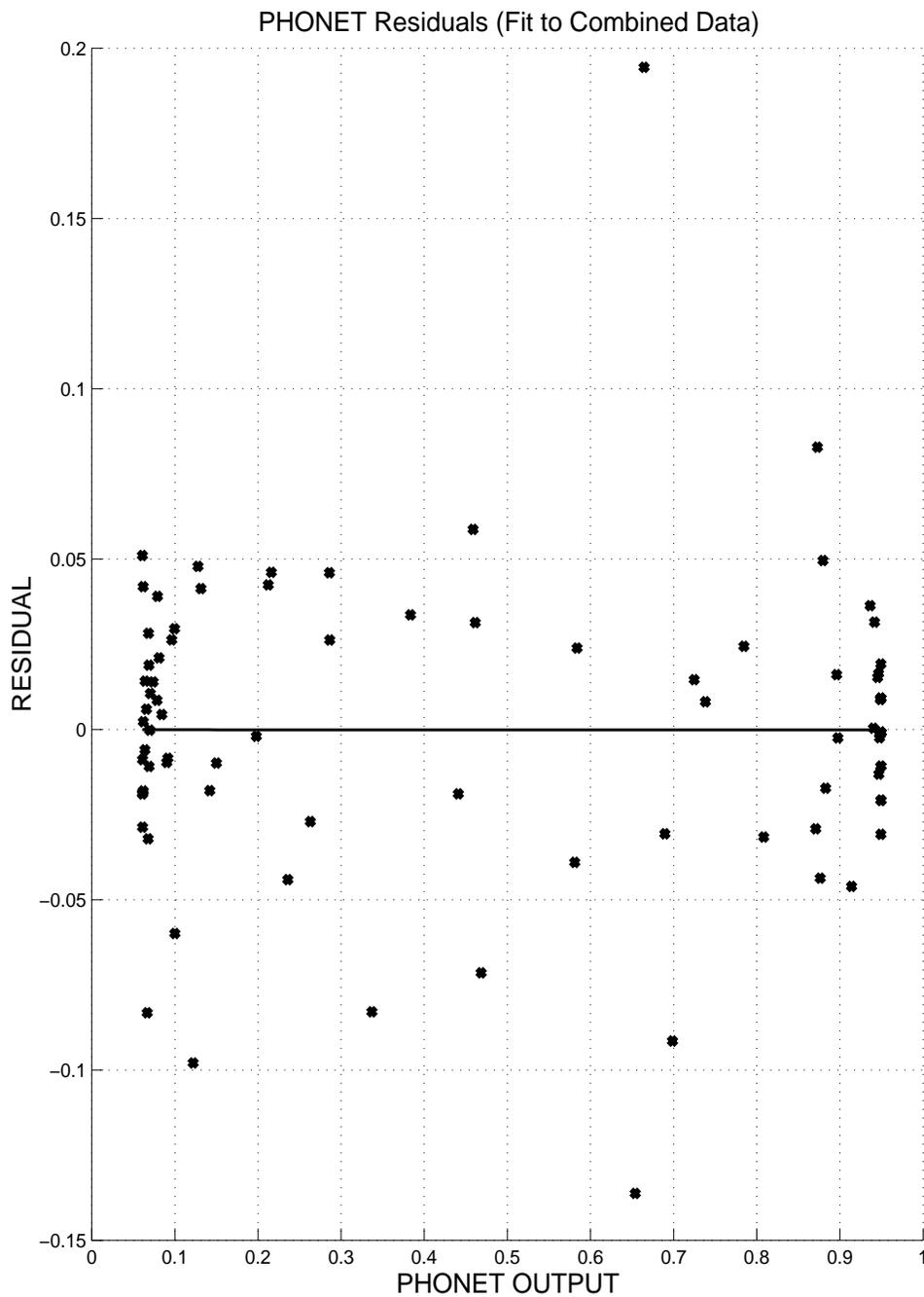
Figure 12: Residuals of the S/T model when fit to the combined data set, plotted against model output. The slope of the linear regression, as well as the expected numbers of positive residuals and runs reveal no trend in the residuals.

mean and the variance of the distributions vary with syllable duration. In addition, Miller and Wayland (1993) and Miller and Volaitis (1994) reported that varying the sentence-level (global) speech rate also shifts the means of the within-category distribution functions, but it does not produce a change in the variance.

Further argument is needed to understand how the brain can defer its classification of information, such as the /b/–/w/ distinction, until it processes later information, such as the duration of a subsequent vowel. One needs also to analyze how this more slowly varying classification process achieves its sensitivity to the S/T ratio. Grossberg (1978, 1986) proposed, as noted in Section 3, that the S/T ratio is represented as a spatial pattern of activation across a working memory, and that this pattern is categorized by a competitive learning or self-organizing feature map network. In brief, the activation pattern generates output signals that are processed by an adaptive filter. The filter generates inputs to a second level of nodes, or cell populations, that categorize the patterns that are active in working memory (Figure 3). A category node, or small set of nodes, is chosen by lateral inhibitory, or competitive, interactions that occur among the category nodes. Only the nodes that receive the largest inputs from the adaptive filter win the competition. Adaptive weights, or long term memory traces, in the filter pathways undergo learning only if they input to a winning node. Learning is designed to encode the ratio of activations across the working memory nodes. This is how category nodes in the model become sensitive to the S/T ratio.

Why does this classification process take so long that the duration of a vowel can influence the percept of the preceding consonant? Why isn't the consonant already classified before the vowel is fully presented? Grossberg (1978, 1986) proposed that the perceptual event is not bottom-up activation of a category node *per se*. Rather, when a category node is activated, it releases learned top-down signals to the working memory. These top-down signals represent a prototype of the chosen category. The prototype is matched against the working memory pattern, and can reorganize it by generating a focus of attention that selects the feature pattern that is expected by the prototype from the total activation pattern.

As this matching process takes hold, it reactivates consistent category nodes via the bottom-up filter. The amplified category nodes, reinforce their top-down prototype signals. This bottom-up and top-down exchange of amplified matching signals generates a resonant state within the system. The resonant state evolves on a slower time scale than bottom-up activation. The resonant state, rather than bottom-up activation *per se*, is assumed to subserve the conscious speech percept. The resonant state is also assumed to trigger any new learning of categories in the bottom-up filter and of prototypes in the top-down expectation. Hence this resonant event has been called an *adaptive resonance*, and the larger theory of which it is a part has been called Adaptive Resonance Theory, or ART (Carpenter and Grossberg, 1991; Grossberg, 1980, 1995).

Within ART, the brain's sensitivity to the S/T ratio is ascribed to the fact that the resonance takes hold slowly enough that the duration of the vowel has a chance to influence the final CV percept. The S/T ratio represents an extension of the old notion of consonant/vowel duration ratio as a cue in phonetic identification. The success of the automatically gain controlled S/T ratio in PHONET's account of context effects in phonetic identification, its function as a stable basis for learning phonetic codes in the larger framework of ART, and the potential generality of the model to other phonetic contrasts hold promise for the application of PHONET to further problems in speech recognition.

Carpenter and Grossberg (1993), Cohen *et al.* (1988), Grossberg (1986), Grossberg *et al.* (1997), and Grossberg and Stone (1986b) have used ART mechanisms to explain a variety of other data about speech and language perception and production. Of particular relevance to the present work are simulation results within this model framework of data concerning how the brain compensates, not just for the local speech rate variations analyzed here, but also for global speech rate variations that take hold on the time scale of an entire utterance (Grossberg *et al.*, 1997). These results were also based on separate automatically gain-controlled working memories for consonants and vowels whose resonant activities contribute to a conscious speech percept. Taken together, these results on local and global speech rate compensation support the hypothesis that gain-controlling interactions at speech working memories help to establish working memory invariants whose categorization via resonant bottom-up and top-down interactions are the basis for conscious perception.

APPENDIX

A   Error bars

The error bars estimated from the original data sets, shown in Figures 4 through 9, were computed from the standard error of the mean for each curve, based on the assumption of a binomial decision process. The variance of a binomial random variable is $\text{Var}\{x\} = \sigma^2 = E\{x_i^2\} - E\{x_i\}^2$, where the expected value is $E\{x_i\} = \sum_i^n x_i f(x_i)$, where $f(x)$ is the probability density. Now each datum shown in Figures 1B and 2B represents the sample mean of subject responses to that stimulus condition. There were $n = 192$ total presentations of each stimulus in the Miller and Liberman experiment and $n = 200$ in the Schwab *et al.* experiment. Let the x_i take on the values 1 or zero (subject replies /b/ or no /b/), so $E\{x_i\} = p \times 1 + (1 - p) \times 0 = p$, where p is the sample probability of /b/. Also, the sum $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i = np$. Then the variance is $\text{Var}\{x\} = \frac{1}{n}np - p^2 = p(1 - p)$. The standard error of the mean, defined as $\sqrt{\text{Var}\{x\}/N}$, then gives us the error in probability for the datum, $\sigma/\sqrt{n} = \sqrt{p(1-p)/n}$.

#	T	(a) χ^2	(a) $100\rho^2$	(b) χ^2	(b) $100\rho^2$
1	$\theta_1 R + \theta_2 E + \theta_3$	216	98.69	133	99.25
2	$\theta_1 R + \theta_2 D + \theta_3$	1917	83.43	155	99.14
3	$\theta_1 E + \theta_2 D + \theta_3$	5789	63.41	181	99.08
4	$\theta_1 ED + \theta_2$	2183	85.65	564	97.78
5	$\theta_1 RD + \theta_2$	8347	14.50	1609	92.51
6	$\theta_1 RE + \theta_2$	2739	80.65	571	99.12
7	$\frac{R+\theta_1}{\theta_2(E+\theta_3)}$	291	98.28	135	99.23
8	$\frac{E+\theta_1}{\theta_2(R+\theta_3)}$	8507	12.49	295	98.35
9	$\frac{R+\theta_1}{\theta_2(D+\theta_3)}$	3448	79.16	222	98.78
10	$\frac{E+\theta_1}{\theta_2(D+\theta_3)}$	2608	85.77	491	97.18
11	$\frac{D+\theta_1}{\theta_2(E+\theta_3)}$	8466	14.24	293	98.30
12	$\frac{D+\theta_1}{\theta_2(R+\theta_3)}$	8497	12.55	298	98.26

Table 4: Goodness of fit obtained using different transient channel expressions.

B   Data fitting procedure

Response probabilities and ratings were obtained from Figure 3 of Miller and Liberman (1979) and Figures 4-6 of Schwab *et al.* (1981). For the Miller and Liberman data ($n = 65$), transition durations varied 16 to 64 msec with frequency extent fixed at 616 Hz, and vowel durations were 80, 116, 152, 224, or 296 msec. For the Schwab *et al.* data ($n = 21$), all

#	$g(R, E)$	χ^2	$100 * \hat{\rho}^2$
1	g_1	3760	70.37
2	$g_1 + g_2 R$	3550	71.38
3	$g_1 + g_2 E$	3550	71.38
4	$g_1 + g_2 RE$	3740	70.07
5	$g_1 + g_2 R + g_3 E$	319	97.65
6	$g_1 + g_2 R + g_3 RE$	2075	88.26
8	$g_1 + g_2 E + g_3 RE$	9662	0
9	$g_1 + g_2 R + g_3 E + g_4 RE$	227	98.53
10	$g_1 + g_2 R + g_3 E + g_4 RE + g_5 E^2$	210	98.57
11	$g_1 + g_2 R + g_3 E + g_4 RE + g_5 R^2$	200	98.67
12	$g_1 + g_2 R + g_3 E + g_4 RE + g_5 E^2 + g_6 R^2$	192	98.64

Table 5: Goodness of fit of different cross-channel gain functions.

syllables had a fixed total duration of 245 msec, so vowel duration was calculated as 245-transition duration. The three F2 transition series were equally spaced with the following endpoint values: Rate constant: duration = 30 to 60 msec, extent = 313 to 626 Hz; Extent constant: duration = 15 to 75 msec, rate = 31.33 to 6.26 Hz/msec; Duration constant: extent = 260 to 680 Hz, rate = 11.33 to 1.17 Hz/msec. To make model parameters commensurate for the purposes of fitting the data, all parameters were linearly rescaled to the range of 16 to 280 units. The optimization routine fmins in Matlab was used to implement the Nelder simplex search for the best parameters.

C   PHONET variants and cross-channel gain functions tested

Twelve variants of PHONET using different expressions for transient channel activity were tested to further examine the roles of rate, extent, and duration. Two and three parameter combinations of R, E, and D were optimized to the combined data using (a) $g(T) = $ constant, and (b) $g(T) = g_1 + g_2 R + g_3 E + g_4 RE$. Table 4 lists the χ^2 and ρ^2 statistics associated with each model.

Twelve versions of the cross-channel gain function $g(T)$ were tested. Beginning with $g(T) = $ constant, linear and quadratic combinations of rate and extent were used to evaluate an appropriate approximation to $g(R, E) \approx g(T) \approx T$, where T is given by Eqn. (3). The models were tested using the combined data set. Resulting χ^2 and $\hat{\rho}^2$ statistics are listed in Table 5. The adjusted proportion of variance $\hat{\rho}^2 = \frac{(N-1)\rho^2 - k}{N - 1 - k}$ where $N = 86 = $ number of data points and $k = $ number of parameters adjusts ρ^2 to account for the better fit expected with more parameters.

References

Assmann, P. F, Nearey, T. M. and Hogan, J. T. (1982) Vowel identification: Orthographic, perceptual and acoustic aspects. *Journal of the Acoustical Society of America*, **71**(4), 975-989.

Baddeley, A.D. (1986). **Working memory.** Oxford: Clarendon Press.

Bailey, P.J. and Summerfield, Q. (1980) Information in speech: Some observations on the perception of s + stop clusters. *Journal of Experimental Psychology: Human Perception and Performance*, **53**(2), 536–563.

Berg, K.M. (1985). Temporal masking level differences for transients: Further evidence for a short-term integrator. *Perception and Psychophysics*, **37**, 397–406.

Bradski, G., Carpenter, G.A., and Grossberg, S. (1992). Working memory networks for learning temporal order with application to three-dimensional visual object recognition. *Neural Computation*, **4**, 270–286.

Bradski, G., Carpenter, G.A., and Grossberg, S. (1994). STORE working memory networks for storage and recall of arbitrary temporal sequences. *Biological Cybernetics*, **71**, 469–480.

Britt, R. and Starr, A. (1976a). Synaptic events and discharge patterns of cochlear nucleus cells. I. Steady-frequency tone bursts. *Journal of Neurophysiology*, **39**, 162–178.

Britt, R. and Starr, A. (1976b). Synaptic events and discharge patterns of cochlear nucleus cells. II. Frequency-modulated tones. *Journal of Neurophysiology*, **39**, 179–194.

Burbeck, S.L. and Luce, R.D. (1982). Evidence from auditory simple reaction times for both change and level detectors. *Perception and Psychophysics*, **32**, 117–133.

Carpenter, G.A. and Grossberg, S. (1991). **Pattern recognition by self-organizing neural networks.** Cambridge, MA: MIT Press.

Carpenter, G.A. and Grossberg, S. (1993). Normal and amnesic learning, recognition, and memory by a neural model of cortico-hippocampal interactions. *Trends in Neuroscience*, **16**, 131–137.

Casseday, J., Ehrlich, D., and Covey, E. (1994). Neural tuning for sound duration: role of inhibitory mechanisms in the inferior colliculus. *Science*, **264**, 847–850.

Cohen, M.A. and Grossberg, S. (1986). Neural dynamics of speech and language coding: Developmental programs, perceptual grouping, and competition for short term memory. *Human Neurobiology*, **5**, 1–22.

Cohen, M.A. and Grossberg, S. (1987). Masking fields: A massively parallel architecture for learning, recognizing, and predicting multiple groupings of patterned data. *Applied Optics*, **26**, 1866–1891.

Cohen, M.A. and Grossberg, S. (1997). Parallel auditory filtering by sustained and transient channels separates coarticulated vowels and consonants. *IEEE Transactions on Speech and Audio Processing*, **5**(4), 301–318.

Cohen, M.A., Grossberg, S., and Stork, D.G. (1988). Speech perception and production by a self-organizing neural network. In Y.C. Lee (Ed.), **Evolution, learning, cognition, and advanced architectures**, Hong Kong: World Scientific Publishers.

Delgutte, B. and Kiang, N.Y.S. (1984a). Speech coding in the auditory nerve I: Vowel-like sounds. *Journal of the Acoustical Society of America*, **75**, 866–878.

Delgutte, B. and Kiang, N.Y.S. (1984b). Speech coding in the auditory nerve II: Processing schemes for vowel-like sounds. *Journal of the Acoustical Society of America*, **75**, 879–886.

Denes, P. (1955). Effects of duration on the perception of voicing. *Journal of the Acoustical Society of America*, bf 27, 761–764.

Derr and Massaro (1980). The contribution of vowel duration F0 contour, and frication duration as cues to the /jus/ – /juz/ distinction. *Perception and Psychophysics*. 51–59.

DeYoe, E.A. and Van Essen, D.C. (1988). Concurrent processing streams in monkey visual cortex. *Transactions in Neuroscience*, **11**(5), 219–226.

Diehl, R.L. and Walsh, M.A. (1989). An auditory basis for the stimulus-length effect in the perception of stops and glides. *Journal of the Acoustical Society of America*, **85**(5), 2154–2164.

Felshiem, C. and Ostwald, J. (1996). Responses to exponential frequency modulations in the rat inferior colliculus. *Hearing Research*, **98**, 137–151.

Grossberg, S. (1973). Contour enhancement, short term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics*, **52**, 217-257. Reprinted in Grossberg, S. (1982). **Studies of Mind and Brain.**

Grossberg, S. (1976). Adaptive pattern classification and universal recoding, I: Parallel development and coding of neural feature detectors. *Biological Cybernetics*, **23**, 121–134.

Grossberg, S. (1978). A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans. In R. Rosen and F. Snell (Eds.), **Progress in theoretical biology, vol. 5.** New York: Academic Press, 233–374. Reprinted in Grossberg, S. (1982). **Studies of Mind and Brain.**

Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review*, **87**, 1–51.

Grossberg, S. (1982). **Studies of Mind and Brain.** Amsterdam: Kluwer Academic Publishers.

Grossberg, S. (1986). The adaptive self-organization of serial order in behavior: Speech, language, and motor control. In E.C. Schwab and H.C. Nusbaum (Eds.), **Pattern recognition by humans and machines, vol. 1: Speech perception.** New York: Academic Press.

Grossberg, S. (1995). The attentive brain. *American Scientist*, **83**, 438–449.

Grossberg, S. (1998). Pitch-based streaming in auditory perception. Technical Report CAS/CNS-TR-96-007, Boston, MA: Boston University. In N. Griffith and P. Todd (Eds.), **Musical networks: Parallel distributed perception and performance**. Cambridge, MA: MIT Press, in press, 1998.

Grossberg, S., Boardman, I., and Cohen, M.A. (1997). Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology: Human Perception and Performance*, **23**(2), 481–503.

Grossberg, S. and Stone, G.O. (1986a). Neural dynamics of attention switching and temporal order information in short term memory. *Memory and Cognition*, **14**(6), 451–468.

Grossberg, S. and Stone, G.O. (1986b). Neural dynamics of word recognition and recall: Attentional priming, learning, and resonance. *Psychological Review*, **93**, 46–74.

Hodgkin, A.L. (1964). **The conduction of the nervous impulse.** Liverpool: Liverpool University.

Hodgkin, A.L. and Huxley, A.F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, **117**, 500–544.

Howell, P. and Rosen, S. (1983). Production and perception of rise time in the voiceless affricate/fricative distinction. *Journal of the Acoustical Society of America*, **73**(3), 976–984.

Irvine, D.R.F., *Progress in Sensory Physiology 7*, Springer-Verlag, Berlin, 1986.

Kohonen, T. (1984). **Self-organization and associative memory.** New York: Springer-Verlag.

Kunisaki, O. and Fujisaki (1977). On the Influence of Context upon Perception of Voiceless Fricative Consonants. *Annual Bulletin, Research Institute of Logopedics and Phoniatrics*, 85–91.

Liberman, A.M., Delattre, P.C., Gerstman, L.J., and Cooper, F.S. (1956). Tempo of frequency change as a cue for distinguishing classes of speech sounds. *Journal of Experimental Psychology*, **52**, 127–137.

Macmillan, N.A. (1971). Detection and recognition of increments and decrements in auditory intensity. *Perception and Psychophysics*, **10**, 233–238.

Macmillan, N.A. (1973). Detection and recognition of intensity changes in tone and noise: The detection-recognition disparity. *Perception and Psychophysics*, **13**, 65–75.

Mann, V. and Repp, B. (1980). The influence of vocalic context on perception of the /sh/-s context. *Perception and Psychophysics*, **28**, 213–228.

Mann, V. and Soli, S.D. (1991). Perceptual order and the effect of vocalic context on fricative perception. *Perception and Psychophysics*, **49**, 399–411.

Massaro, D.W. and Cohen, M.M. (1993). The paradigm and the fuzzy logical model of perception are alive and well. *Journal of Experimental Psychology: General*, **122**, 115–124.

Mendelson, J., Schreiner, C., Sutter, M, and Grasse, K. (1993). Functional topography of cat primary auditory cortex: responses to frequency-modulated sweeps. *Experimental Brain Research*, **94**, 65–87.

Miller, J.L. (1981). Effects of speaking rate on segmental distinctions. In P.D. Eimas and J.L. Miller (Eds.), **Perspectives on the study of speech.** Hillsdale, NJ: Lawrence Erlbaum Associates.

Miller, J.L. (1987). Effects of speaking rate on segmental distinctions. In A.D. Ellis (Ed.), **Progress in the psychology of language, vol. 3**. Hillsdale, NJ: Lawrence Erlbaum Associates.

Miller, J.L. and Liberman, A.M. (1979). Some effect of later-occurring information on the perception of stop consonant and semivowel. *Perception and Psychophysics*, **25**(6), 457–465.

Miller, J.L. and Volaitis, L.E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception and Psychophysics*, **46**(6), 505–512.

Miller, J.L. and Wayland, S.C. (1993) Limits on the limitations of context–conditioned effects in the perception of (b) and (w). *Perception and Psychophysics*, **54**, 205–210.

Møller, A.R. (1983). **Auditory physiology**. New York: Academic Press.

Nabet, B. and Pinter, R. (1991). **Sensory neural networks: lateral inhibition**. Boca Raton, FL: CRC Press.

Newman, R., and Sawusch, J. (1996). Perceptual normalization for speaking rate: Effects of temporal distance. *Perception and Psychophysics*, **58**(4), 540–560.

Nittrouer, S. and Studdert-Kennedy, M. (1986). The stop-glide distinction: Acoustic analysis and perceptual effect of variation in syllable amplitude envelope for initial /b/ and /w/. *Journal of the Acoustical Society of America*, **80**(4), 1026–1029.

Phillips, D.P. (1993). Neural representation of stimulus times in the primary auditory cortex. *Annals of the New York Academy of Sciences*, **682**, 104–119.

Pickles, J.O. (1988). **An introduction to the physiology of hearing, 2nd edition**. San Diego: Academic Press.

Pisoni, D.A., Carrell, T.D., and Gans, S.J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception and Psychophysics*, **34**, 314–322.

Port, R. F. and Dalby, J. (1982). Consonant/vowel ratio as a cue for voicing in English. *Perception and Psychophysics*, **32**, 315–322.

Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T. (1988). **Numerical recipes in C: The art of scientific computing**. Cambridge, England: Cambridge University Press.

Rao, C. R. (1973). **Linear statistical inference and its applications.** Wiley: New York.

Repp, B. (1978). Perceptual integration and differentiation of spectral cues for intervocalic stop consonants. *Perception and Psychophysics*, **24**(5), 471–485.

Repp, B. and Liberman, A. (1987). Phonetic category boundaries are flexible. In Harnad, S. N. (Ed.), **Categorical Perception: The Groundwork of Cognition**, chap. 3, pp. 89–112. Cambridge, University Press, New York.

Rhode, W.S. and Smith, P.H. (1986). Physiological studies on neurons in the dorsal cochlear nucleus of the cat. *Journal of Neurophysiology*, **56**, 287–307.

Rumelhart, D.E. and Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive Science*, **9**, 75–112.

Sachs, M.B. and Young, E.D. (1979). Encoding of steady state vowels in the auditory nerve: Representations in terms of discharge rate. *Journal of the Acoustical Society of America*, **66**, 470–479.

Scheffe, H. (1959). **The analysis of variance.** Wiley: New York.

Schwab, E.C., Sawusch, J.R., and Nusbaum, H.C. (1981). The role of second formant transitions in the stop-semivowel distinction. *Perception and Psychophysics*, **21**, 121–128.

Shamma, S.A., Versnel, H., and Kowalski, N. (1995). Ripple analysis in ferret primary auditory cortex. I. Response characteristics of single units to sinusoidally rippled spectra. *Auditory Neuroscience*, **1**, 233–254.

Shamma, S.A. and Symmes, D. (1985). Patterns of inhibition in auditory cortical cells in awake squirrel monkeys. *Hearing Research*, **19**, 1–13.

Shinn, P. and Blumstein, S. (1984). On the role of the amplitude envelope for the perception of [b] and [w]. *Journal of the Acoustical Society of America*, **75**(4), 1243–1252.

Shinn, P., Blumstein, S., and Jongman (l985). The limitations of context conditioned effects in the perception of [b] and [w]. *Perception and Psychophysics*, **38**, 397–407.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, **7**(5), 1074–1095.

Tian, B. and Raushecker, J. (1994). Processing of frequency-modulated sounds in the cat's anterior auditory field. *Journal of Neurophysiology*, **71**(5), 1959–1975.

van Wieringen, A. and Plos, L. (1995a). Frequency and duration discrimination of short first-formant speechlike transitions. *Journal of the Acoustical Society of America*, **95**(1), 502–511.

van Wieringen, A. and Plos, L. (1995b). Discrimination of single and complex consonant–vowel and vowel–consonant-like formant transitions. *Journal of the Acoustical Society of America*, **98**(3), 1304–1312.

Volaitis, L.E. and Miller, J.L. (1992). Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *Journal of the Acoustical Society of America*, **92**, 723–735.

Walsh, M. A. and Diehl, R. L. (1991). Formant transition duration and amplitude rise times as cues to the Stop/Glide distinction. *Quarterly Journal of Experimental Psychology*, **43**(3), 603–620.

Young, E.D. and Sachs, M.B. (1979). Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory nerve fibers. *Journal of the Acoustical Society of America*, **66**, 1381–1403.

Zurita, P., Villa, A., de Ribbaupierre, Y., and Rouiller, E. (1994). Changes of single unit activity in the cat's auditory thalamus and cortex associated to different anesthetic conditions. *Neuroscience Research*, **19**, 303–316.